Brazilian Journal of Probability and Statistics
2017, Vol. 31, No. 4, 765–800
https://doi.org/10.1214/17-BJPS375
© Brazilian Statistical Association, 2017

Studying the effective brain connectivity using multiregression dynamic models

Lilia Costa[a], Thomas Nichols[b,c,d] and Jim Q. Smith[e]

[a]*Universidade Federal da Bahia*
[b]*Oxford Big Data Institute*
[c]*Li Ka Shing Centre for Health Information*
[d]*Discovery, Nuffield Department of Population Health, University of Oxford*
[e]*The University of Warwick*

Abstract. The *Multiregression Dynamic Model* (MDM) is a multivariate graphical model for a multidimensional time series that allows the estimation of time-varying effective connectivity. An MDM is a state space model where connection weights reflect the contemporaneous interactions between brain regions. Because the marginal likelihood has a closed form, model selection across a large number of potential connectivity networks is easy to perform. With application of the Integer Programming Algorithm, we can quickly find optimal models that satisfy acyclic graph constraints and, due to a factorisation of the marginal likelihood, the search over all possible directed (acyclic or cyclic) graphical structures is even faster. These methods are illustrated using recent resting-state and steady-state task fMRI data.

1 Introduction

In this paper, we estimate the effective connectivity for resting-state and steady-state task-based functional Magnetic Resonance Imaging (fMRI) data, using a class of Dynamic Bayesian Network (DBN) models, called the Multiregression Dynamic Model (MDM). The MDM is a model for multivariate time series and aims to efficiently identify dynamic dependence across variables over time (Queen and Smith, 1993, Queen and Albers, 2009). In the study of *functional integration*, which considers how different parts of the brain work together to give rise to behavior and cognition, a distinction is made between *functional connectivity* and *effective connectivity* (Poldrack, Mumford and Nichols, 2011). The former is defined as the correlation among measurements of neuronal activity of different areas whilst the effective connectivity represents causal influence from one brain region onto another (Friston, 2011). Several models have been developed in order to define and detect a causal flow from one variable to other, especially in the area of machine learning (e.g., Spirtes, Glymour and Scheines, 2000 and Pearl, 2000). The MDMs embody a particular pattern of causal relationships which, unlike the

Key words and phrases. Multiregression dynamic model, Bayesian network, effective connectivity, functional magnetic resonance imaging, integer programming algorithm.
Received September 2016; accepted August 2017.

Bayesian Network (BN), expresses the *dynamic* message passing as well as the potential connectivity between different areas in the brain.

There are different methods in neuroscience literature which aim to estimate functional and effective connectivity. They usually associate a model to a graphical structure. A graph consists of *nodes* and *edges*, where the latter represents connections between pairs of the former. In connectivity studies, nodes are typically considered to be the regions of interests (ROI's), and the goal of each method is to detect the edges that connect the regions (Sporns, 2010). When one node influences other, the edge indicates the direction of the effect. In the example shown in Figure 1, the edges are *directed* and, as no path starts and ends at the same node, they are called *directed acyclic graphs* (DAGs). When there is a directed edge from one node to another, the former is called a *parent* while the latter is a *child*.

One of the most popular methods for modeling brain dependences in neuroimaging is the *Dynamic causal modeling* (DCM). The DCM estimates the directed effective connectivity through a model formulated with stochastic differential equations (Stephan et al., 2008). The most common DCM is a deterministic one, where it is assumed that the neuronal activity is completely determined by the model, and so it is not intended to be used for resting-state connectivity (Penny, Ghahramani and Friston, 2005, Smith et al., 2011). The DCM uses a detailed biophysical model to connect the neuronal activity to the measurable fMRI data, making the inference process quite complex and infeasible for more than just a few nodes (Friston, Harrison and Penny, 2003, Poldrack, Mumford and Nichols, 2011).

Another popular model is the *Bayesian Network* which is a simpler and non-dynamic model that expresses causal concepts in terms of conditional independence among variables (Smith and Croft, 2003, Goldenberg et al., 2010). A BN is a specific case of the MDM when connections are constant over time. Other classes of dynamic linear models have been developed recently, such as the *Linear Dynamic System* (LDS; Smith et al., 2010, 2011) and the *Multivariate* (or Bilinear) *Dynamical System* (BDS; Penny, Ghahramani and Friston, 2005, Ryali et al., 2011). While they are more sophisticated than the MDM, these other models usually estimate static connectivity and their scores are not factorable, and hence are much slower to search over.

In an MDM, the directed edges can be associated with a potential causal directionality, as argued in Queen and Albers (2009), and hint at the *effective* connectivity rather than the *functional* connectivity. The MDMs are a class of directed graphical models with a number of appealing features: (i) In contrast to standard Bayesian Networks, an MDM explicitly models the changes in the connectivity over time; (ii) Critically, the MDM is driven by contemporaneous interactions between the brain regions rather than lagged relationships as some methods used, such as Dynamic Bayesian Network (Goldenberg et al., 2010) and methods based on Granger causality (see below); (iii) The MDMs need not be acyclic, though when constrained to be acyclic they can dissociate otherwise indistinguishable

Markov-equivalent graphs (see Section 4); (iv) But perhaps the biggest advantage of the MDM as compared to other dynamic models is their closed form for the marginal likelihood. The marginal likelihood is used to score models and estimate parameters, and can be written as a product of multivariate Student's t distributions conditional on a small number of smoothing parameters (details below). This closed form allows us to perform model selection over a dynamic class of models very quickly, just as for the simple BN; (v) In addition, if it is necessary to include other features, like change points or covariates, it is straightforward to embellish the MDM. Even after that, the MDM often exhibits a closed form likelihood distribution (Costa et al., 2015). This means that it is possible to search over many candidate networks quickly, estimating each of their parameters and hence their time profile of changing strengths along the way.

One particular embellishment is to use lagged information, that is, to add the past of variables as parents, creating a Granger causality MDM (Granger, 1969, Havlicek et al., 2010). Classes like this one that directly model Granger causality have received severe criticism when applied to the fMRI datasets (Chang, Thomason and Glover, 2008, David et al., 2008, Valdés-Sosa et al., 2011, Smith et al., 2012). For instance, the match between the sampling interval and the time constant in the neurodynamics is often poor, because the temporal delay blood-based response can vary considerable across brain regions. In fact, (Smith et al., 2011) discovered that lag-based approaches like these do not perform well at identifying connections for fMRI data, albeit only under the assumption of static connectivity strength. Therefore we do not consider this extension further.

In this paper, we use two search methods, one for acyclic graphs, and another for directed (acylic or cyclic) graphs. We call the first the MDM-IPA, which uses the Integer Programming Algorithm to search for the acyclic graphical structures. This method has showed good performance for synthetic and real resting-state fMRI data; in particular, we have shown that it provides comparable performance to methods like Patel's tau and generalized synchronization (Costa et al., 2015). The other search method is called the MDM-DGM, which does not consider the acyclic constraints and searches the larger class of directed graphs and we present here for the first time.

The purpose of this paper is to present the development of this new search method the MDM-DGM and a comparison with the MDM-IPA. The MDM-DGM appears to be more suitable for neuroscience because unlike the MDM-IPA it is able to model a bidirectional communication between brain regions which typically exists in this domain. In this paper, we also present the first application of the MDM-DGM to two different fMRI studies (resting-state and task fMRI data) and the first application of the MDM-IPA to task fMRI data. The first application consists of a resting-state experiment acquired on 25 subjects and 3 sessions per subject, considering 4 brain regions, and aims to study the information flow of the brain, *for example*, "forward" or "backward". The second one provides data from 5 different experiments, considering 11 brain regions and 15 subjects, and

aims to compare the map of brain connections for different conditions, such as resting-state and motor activity.

The remainder of this paper is structured as follows. Section 2 describes the MDM while Section 3 provides search network algorithms. In Sections 4 and 5, these search methods are applied into resting-state and steady-state task real datasets, respectively. Finally, Section 6 concludes the work and describes the directions for future work.

2 The multiregression dynamic model

The Multiregression Dynamic Model (MDM) is a multivariate model for an observable series which is broken down into univariate regression dynamic linear models (DLMs; West and Harrison, 1997), one for each node. These DLMs allow the effective connectivity to vary during the period of investigation. To introduce the DLM, consider the relationship between 3 regions: Posterior Cingulate, Anterior Frontal and Left Lateral Parietal, with their respective time series being $Y_t(1)$, $Y_t(2)$ and $Y_t(3)$, representing the averaging time series over all voxels within the region. A possible model linking one of the series to the two others might be

$$Y_t(1) = \theta_t^{(1)} + \theta_t^{(2)} Y_t(2) + \theta_t^{(3)} Y_t(3) + v_t,$$

where $\theta_t^{(1)}$ is the baseline, $\theta_t^{(2)}$ and $\theta_t^{(3)}$ are the coefficients for regions 2 and 3, respectively, and v_t is an error term. The parameters $\theta_t^{(2)}$ and $\theta_t^{(3)}$ are defined as effective connectivity and can be seen as a temporal dependence between brain areas and therefore it may be defined as dynamic (activity-dependent). Thus these parameters are all time-dependent, allowing the influence of Anterior Frontal and Left Lateral Parietal into Posterior Cingulate to vary over time. Considering a walk-random model, the regression parameters are written in function of their past values as

$$\theta_t^{(j)} = \theta_{t-1}^{(j)} + w_t^{(j)},$$

where $w_t^{(j)}$ is innovation, for $j = 1, \ldots, 3$. It is interesting to note that other dynamic models of this environment, for example Dynamic Causal Modeling (Friston, Harrison and Penny, 2003, Stephan et al., 2008), albeit having technical differences in definition, are similarly modeled as a function of latent variables which follow a Markovian dynamic process. Furthermore, unlike non-dynamic directed acyclic graph (DAG), Markov equivalent DAGs have different associated predictive distributions and thus can be distinguished. Thus, at least in principal, it is possible to use the MDM to discriminate patterns of causal directionality, that would be undetectable with BN models (Queen and Albers, 2009, Costa et al., 2015), see an example of this in Section 4.

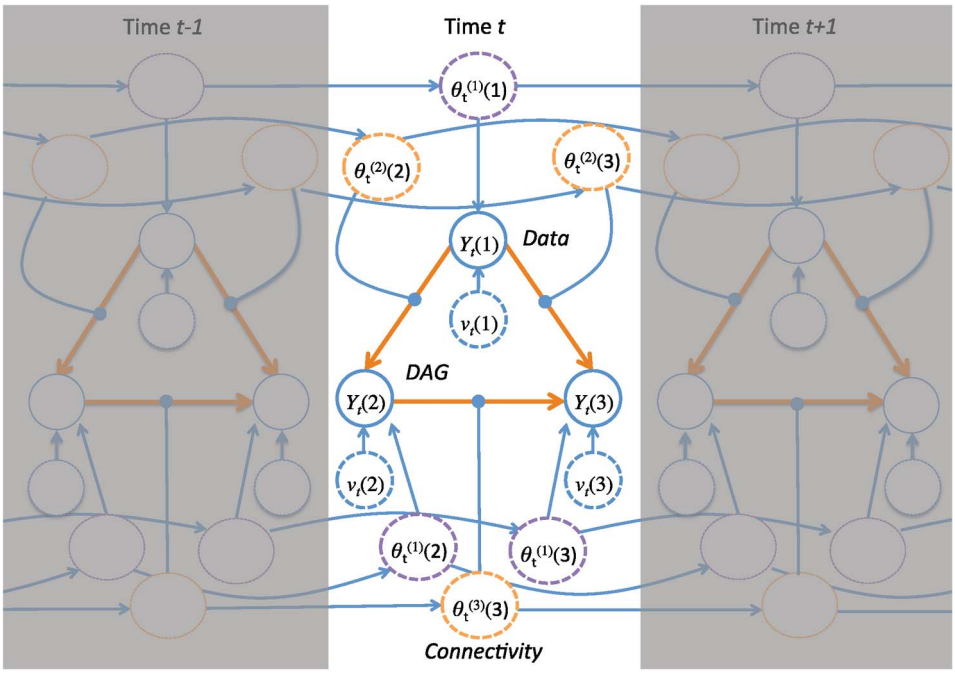

Figure 1 *Dependence structure for the MDM considering region* 1 *as the parent of region* 2 *and region* 3; *and region* 2 *as the parent of region* 3. *The solid circles represent observed variables*, $Y_t(r)$, $r = 1, 2, 3$. *The dashed circles represent latent variables*: *blue for observational errors*, $v_t(r)$; *violet for the intercept of the regression of region* r, $\theta_t^{(1)}(r)$; *and orange for the effective connectivity strength between two regions*, $\theta_t^{(2)}(2)$, $\theta_t^{(2)}(3)$ *and* $\theta_t^{(3)}(3)$.

The description of the model

Before providing a detailed description of the MDM, we will introduce some notation. Let $\mathbf{Y}_t' = (Y_t(1), \ldots, Y_t(n))$ denote the variable at time t for the n regions. Their observed values are designated respectively by $\mathbf{y}_t' = (y_t(1), \ldots, y_t(n))$. We define $\mathrm{Pa}(r)$ to be the set of parents for region r, and denote $\mathbf{Y}^t(r)' = (Y_1(r), \ldots, Y_t(r))$ as the previous information up to time t for region $r = 1, \ldots, n$.

The linear multiregression dynamic model (LMDM) is defined by n observation equations, system equation and initial information (Queen and Smith, 1993). A DAG representing an MDM for three nodes is shown in Figure 1, and in full generality is as follows.

Observation equations:

$$Y_t(r) = \mathbf{F}_t(r)'\boldsymbol{\theta}_t(r) + v_t(r), \qquad v_t(r) \sim \mathcal{N}\big(0, V(r)\big);$$

System equation:

$$\boldsymbol{\theta}_t = \mathbf{G}_t\boldsymbol{\theta}_{t-1} + \mathbf{w}_t, \qquad \mathbf{w}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{W}_t); \tag{2.1}$$

Initial information:

$$(\boldsymbol{\theta}_0|y_0) \sim \mathcal{N}(\mathbf{m}_0, \mathbf{C}_0); \tag{2.2}$$

$$(\phi(r)|y_0) \sim \mathcal{G}\left(\frac{n_0(r)}{2}, \frac{d_0(r)}{2}\right), \tag{2.3}$$

where $r = 1, \ldots, n$; $t = 1, \ldots, T$; $\mathcal{N}(\cdot, \cdot)$ is a Gaussian distribution and $\mathbf{F}_t(r)$ is a known function of $\mathrm{Pa}(r)$. In an MDM, $\mathbf{F}_t(r)$ is usually defined as $\mathbf{F}_t(r) = \mathbf{M}(r)\mathbf{Y}_t^*$, where $\mathbf{M}(r)$ is $p_r \times (n+1)$ matrix containing only zeros and ones, where ones indicate the parents of $Y_t(r)$ and the first row of $\mathbf{M}(r)$ is $(1, 0, \ldots, 0)$ representing the intercept; $p_r = |\mathrm{Pa}(r)| + 1$ counts the number of parents of r plus one for the intercept of region r; $\mathbf{Y}_t^* = (1, Y_t(1), \ldots, Y_t(n))'$. The time-varying regression coefficients are $\boldsymbol{\theta}_t' = (\boldsymbol{\theta}_t(1)', \ldots, \boldsymbol{\theta}_t(n)')$, where $\boldsymbol{\theta}_t(r)' = (\theta_t^{(1)}(r), \ldots, \theta_t^{(p_r)}(r))$ is the p_r-dimensional state vector for region r. The parameter $\theta_t^{(1)}(r)$ represents the intercept of the regression of region r whilst $\theta_t^{(i)}(r)$ for $i > 1$ represents the effective connectivity strength for the $(i-1)$th parent of region r. The parameters at time t depend on time $t-1$ through $\mathbf{G}_t = \mathrm{blockdiag}\{\mathbf{G}_t(1), \ldots, \mathbf{G}_t(n)\}$, where $\mathbf{G}_t(r)$ is a $p_r \times p_r$ matrix. While we have included $\mathbf{G}_t$ to indicate the possible enrichment of the system model, in practice this work only considers $\mathbf{G}_t(r) = \mathbf{I}_{p_r}$, where $\mathbf{I}_{p_r}$ is p_r-dimensional identity matrix, and hence $\mathbf{G}_t$ is omitted from now on. This is the simplest choice of $\mathbf{G}_t(r)$ and defines a random walk process on the connectivity strength which appears to fit this application very well. However, in some applications, it can be desirable to define other $\mathbf{G}_t(r)$. For example, if the assumption that the standardized conditional one-step forecast errors are serially independent is not verified and it is necessary to include the ARMA components in DLM form, West and Harrison (1997, Chapter 9) suggest that $\mathbf{G}_t(r)$ is defined as a function of other parameters. However, in this case, the estimation process becomes more complex and much less compatible with the efficient model search methods we use here.

The error terms of this model are $v_t(r)$ and $\mathbf{w}_t$. The observational error, $v_t(r)$, is taken to be independent over t, with variance $V(r)$. The $\mathbf{w}_t$ are innovations for the latent regression coefficients with covariance $\mathbf{W}_t = \mathrm{blockdiag}\{\mathbf{W}_t(1), \ldots, \mathbf{W}_t(n)\}$, each $\mathbf{W}_t(r)$ being a $p_r \times p_r$ matrix. Note that when $\mathbf{W}_t = \mathbf{0}$ for all t, the usual static regression model is obtained. These two terms account for two different types of variance. The $\mathbf{w}_t$ models the slowly varying behavior expected in the $\boldsymbol{\theta}_t$'s, while $v_t(r)$ accounts for uncorrelated noise not otherwise modelled. For more details of this model class see Queen and Smith (1993) and Costa et al. (2015).

When the observational variances are unknown, we write the observation precision as $\phi(r) = V(r)^{-1}$ and the prior information is provided through the distribution of $\boldsymbol{\theta}_0$ and $\phi(r)$, which specifies the information at time $t = 0$, as shown in equations (2.2) and (2.3). The model may be conveniently reparameterised as $\mathbf{W}_t(r) = V(r)\mathbf{W}_t^*(r)$ and the posterior variance is $\mathbf{C}_t(r) = V(r)\mathbf{C}_t^*(r)$. When

$\mathbf{W}_t^*(r)$ is unknown, the state innovation variance can be defined indirectly in terms of a single *discount factor* δ(r). That is, $\mathbf{W}_t^*(r)$ can be defined deterministically through δ(r) as a "discounted" value of $\mathbf{C}_{t-1}^*(r)$. As details in Appendix A, the discount factor can be estimated simply: Since the full posteriors have a closed form conditional on $\delta(r)$, we can find $\delta(r)$ that maximizes the predictive likelihood described below. By doing things this way, the marginal likelihood is of closed form so hugely more efficient to calculate. This is what makes the search over this elaborate model space feasible. For the exploratory phase of the process, it is necessary to take some short cuts like these. We note however that through sensitivity analyses we have been able to check that the marginal likelihood scores of different models are not highly sensitive to these settings. So results associated with the model selection should not be that different to a full Bayesian analysis with properly set up priors. And we have found that this so analysis is often sufficient to see the data against expert judgments from the client to better understand what lies in it. Enough of their domain knowledge is embedded in the closed form analysis for them to properly interpret the outputs.

Of course once a model (or class of high scoring models) has been selected by the scientist for much closer scrutiny then we would advise more principled estimation. But this would often include other features like change points and non-linear dynamics which the scientist might believe also exist which we have ignored in the initial exploratory phase for computational reasons. Indeed the accommodation of these sorts of features often has a much great impact on the output than estimation of these hyperparameters.

Because the equations of the LMDM can be viewed as a collection of nested univariate DLMs, the parameters can be estimated using well-known Kalman Filter recurrences over time. In particular, these standard techniques show that the posterior filtered distributions are written as

$$\big(\boldsymbol{\theta}_t(r)|\mathbf{y}^t\big) \sim \mathcal{T}_{n_t(r)}\big(\mathbf{m}_t(r), \mathbf{C}_t(r)\big) \quad \text{and}$$
$$\big(\phi(r)|\mathbf{y}^t\big) \sim \mathcal{G}\left(\frac{n_t(r)}{2}, \frac{d_t(r)}{2}\right),$$

where $\mathcal{T}_{n_t(r)}(\cdot,\cdot)$ is a Noncentral t distribution with $n_t(r)$ degrees of freedom, and these parameters are easily found through a recurrence relations (see Appendix A). The "filtering estimate" considers data until the present moment, *i.e.* $\mathbf{Y}^t$, and it is used in the calculation of the predictive likelihood and, at the final time point, for model selection. However it is also possible to create retrospective estimates using the entire time series, leading to a so-called *smoothing estimate*. The smoothing estimates are obtained based on the complete dataset, *that is*, $\mathbf{Y}^T$, and so tend to have smaller variance, as shown in Section 4.

The conditional forecast distribution of $(Y_t(r)|\mathbf{y}^{t-1}, \mathrm{Pa}(r))$ is given by:

$$\big(Y_t(r)|\mathbf{y}^{t-1}, \mathrm{Pa}(r)\big) \sim \mathcal{T}_{n_{t-1}(r)}\big(f_t(r), Q_t(r)\big), \tag{2.4}$$

where, again, the parameters are directly computed (see Appendix A). To whom it wants to run the MDM, there is an R package in Schwab et al. (2017) and a R code in the supplementary material of this paper.

Criteria for model selection

One of the most popular ways of comparing two models is to use a Bayes factor (Jeffreys, 1961, West and Harrison, 1997). This is defined as the ratio between the predictive likelihood of two models, model 0 and model 1, say. From the conditional forecast distribution (equation (2.4)), the joint log predictive likelihood (LPL) can be calculated as

$$\mathrm{LPL}(m) = \sum_{r=1}^{n} \sum_{t=1}^{T} \log p\big(Y_t(r) | \mathbf{y}^{t-1}, \mathrm{Pa}(r)\big), \tag{2.5}$$

where m denotes the current choice of model that determines the relationship between the n regions. To compare model m_1 to model m_0, we use a Bayes factor (BF) so that, on the log scale,

$$\log(\mathrm{BF}) = \mathrm{LPL}(m_1) - \mathrm{LPL}(m_0).$$

Priors

As shown above, the prior information is provided through the distribution of $\boldsymbol{\theta}_0$ which specifies the information at time $t = 0$. Here we will show the impact of the parameter priors on the inference process. Consider, for example, node 1 is the only parent of node 2. Under the observation equation: $Y_t(2) = \theta_t Y_t(1) + \upsilon_t(2)$, the conditional forecast mean of the variable $Y_{t+1}(2)$ is $m_t(2)Y_{t+1}(1)$, as in equation (A.8). Furthermore from equation (A.4), $m_t(2)$ can be rewritten as

$$m_t(2) = A_t(2)Y_t(2) + m_{t-1}(2)\big(1 - A_t(2)Y_t(1)\big),$$
$$\text{where } A_t(2) = \frac{R_t^*(2)Y_t(1)}{R_t^*(2)Y_t(1)^2 + 1}.$$

Thus, the calculation of the current forecast mean is based on $(1 - A_t(2)Y_t(1))\%$ of the previous mean m_{t-1}. When the latter is replaced by its own equation in the function of $m_{t-2}(2)$, we find that

$$\begin{aligned} m_t(2) = {} & A_t(2)Y_t(2) + A_{t-1}(2)Y_{t-1}(2)\big(1 - A_t(2)Y_t(1)\big) \\ & + m_{t-2}(2)\big(1 - A_{t-1}(2)Y_{t-1}(1)\big)\big(1 - A_t(2)Y_t(1)\big). \end{aligned}$$

The forecast mean of the second previous time m_{t-2} contributes $(1 - A_{t-1}(2)Y_{t-1}(1))(1 - A_t(2)Y_t(1))\%$ to current forecast mean. Following the

same reasoning for $m_{t-2}(2)$ onwards, we find the forecast mean can be expressed as a function of the prior mean as

$$m_t(4) = A_t(2)Y_t(2) + \sum_{k=1}^{t-1}\left[A_k(2)Y_k(2)\prod_{j=k+1}^{t}\big(1 - A_j(2)Y_j(1)\big)\right] + m_0(2)\prod_{i=1}^{t}\big(1 - A_i(2)Y_i(1)\big). \tag{2.6}$$

Note that as t increases the value of $\prod_{i=1}^{t}(1 - A_i(2)y_i(1))$ therefore decays to zero. So the importance of the prior mean $m_0(2)$ in the calculation of $m_t(2)$ decreases as t increases.

We studied the impact of the prior distribution in the calculation of the posterior distribution of a regression parameter using real fMRI data, with 176 time points and 36 subjects. This inferential process was led using the values of 0 and 1 for the hyperparameter $m_0(2)$ and the values of 0.5, 1 and 3 for the hyperparameter $C_0^*(2)$. Figure 2 (*left*) shows the average of the contribution of prior mean $m_0(2)$ in the calculation of posterior mean $m_t(2)$ (as in the equation (2.6)) over 36 subjects. Note that this contribution is less than 1% after time 17 for all values of prior hyperparameters (a similar result can be seen for the constant model in West and Harrison, 1997, Chapter 2). The central graph shows the posterior mean $m_t(2)$ and the right hand graph shows the posterior variance $C_t(2)$ for a particular subject with the same values of hyperparameters. In general, the average of difference between the results of the prior hyperparameters over subjects is less than 0.02 for the posterior mean from time 11 and less than 0.002 for the posterior variance from time 12. Therefore, after time 10 the posterior distribution is almost the same regardless of the typical values we might choose for the hyperparameters of the prior (see, in the central and right hand graphs, that the different colour lines become almost the only one after the point 10).

3 Searching the MDM using search-and-score methods

Estimating a graphical structure is a formidable problem, since the number of possible causal models grows exponential with the number of nodes considered. For instance, there are over 1 billion possible graphical structures for a BN with just 7 nodes (Sloane and Plouffe, 1995). Several search algorithms have been developed recently to learn BN structure (see, *e.g.*, Spirtes, Glymour and Scheines, 2000, Ramsey et al., 2010, Cussens, 2010, Cowell, 2013). Here we apply an Integer Programming (IP) algorithm to search for DAGs, and a modification of this method to search a Directed Graph Model (DGM).

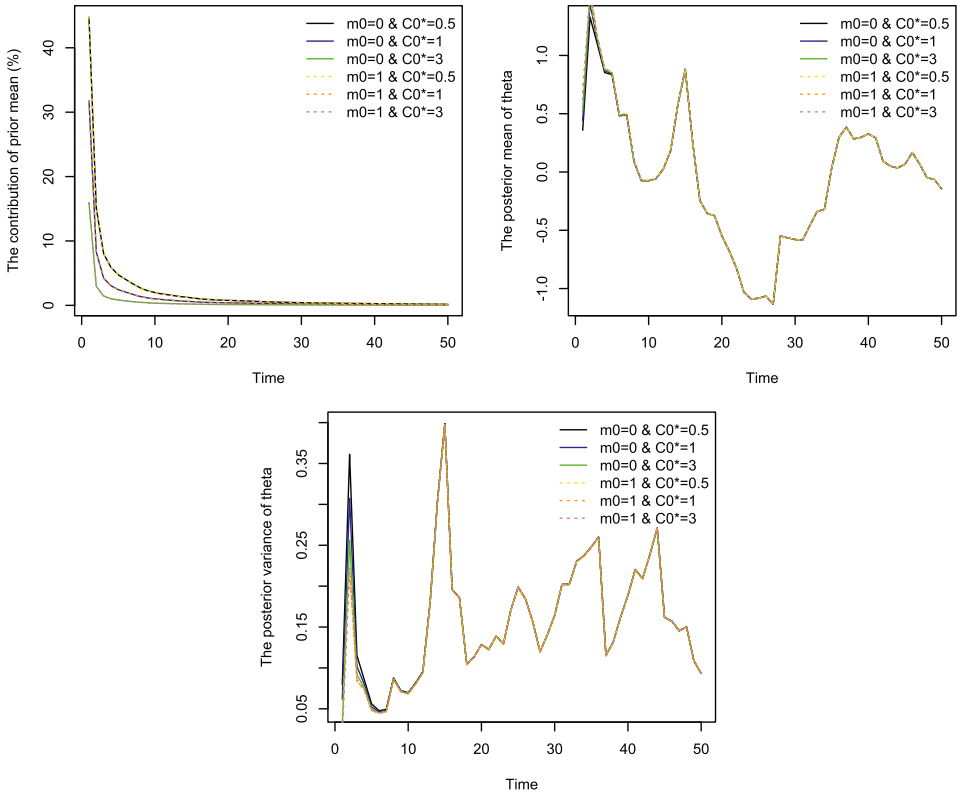

Figure 2 *In this picture we show the impact of priors in the posterior distribution of the connectivity* $Y(1) \to Y(2)$. *The left-hand graph shows the average of the contribution of the prior mean* $m_0(2)$ *in the calculation of posterior mean* $m_t(2)$, *defined as* $\prod_{i=1}^{t}(1 - A_i(2)y_i(2))\%$, *over* 36 *subjects, by different values of hyperparameters. They are less than* 1% *from time* 17. *The central graph shows the posterior mean* $m_t(2)$ *whilst the right hand graph shows the posterior variance* $C_t(2)$ *for a particular subject by the same values of hyperparameters. See text for more details.*

Integer programming algorithm

An Integer Programming algorithm is a search-and-score method, we use to estimate network structure with the MDM scores (MDM-IPA). A *standard* form of IP is defined as the problem of maximizing $\mathbf{c}'\mathbf{x}$, such that $\mathbf{x}$ is integer and both $\mathbf{Ax} \leq \mathbf{b}$ and $\mathbf{x} \geq \mathbf{0}$ are satisfied, given $\mathbf{c}$, $\mathbf{A}$ and $\mathbf{b}$ (Williams, 2009). In our case, from equation (2.5), $\mathrm{LPL}(m)$ is written as a sum of n local scores, say $c(r, \mathrm{Pa}_m(r))$, one for each node r, so that $\mathrm{LPL}(m) = \sum_{r=1}^{n} c(r, \mathrm{Pa}_m(r))$. Therefore, the choice of parent set $\mathrm{Pa}_m(r)$ specified by the candidate model m determines the local score for $Y_t(r)$, and a model selection for MDM can be seen as a search for n sets of parents, $\mathrm{Pa}(1), \ldots, \mathrm{Pa}(n)$, that maximise the LPL, subject to this configuration of parents corresponding to a valid MDM. In the standard IP form, this problem has $n2^{n-1}$ unknowns, with $\mathbf{x}$ being a binary vector, indicating, for each of n nodes, which

of the 2^{n-1} possible sets of parents are active; **c** corresponds to all possible local scores; and **A** and **b** express constraints that ensure **x** corresponds to a valid MDM. There are 2 types of constraints, *n convexity constraints* that ensure each region has exactly 1 set of parents, and *cluster constraints* that ensure the solution is a DAG. The cluster constraints assert that for any subset of nodes in a graph, there must be at least one node with no parents in that subset. We solve this IP problem with the *gobnilp* system (Cussens, 2012, Bartlett and Cussens, 2013) which uses the SCIP IP framework (Achterberg, 2007).

Directed graph model

Bidirectional communication between some brain regions is often expected. Thus cyclic graphs (*i.e.*, graphs allowing cycles) may better represent brain networks than DAGs. Therefore, we are also considering search for graphical structure without the constraints of DAG (*cluster constraints*). Because the predictive likelihood factors by node (modularity property), optimisation can be done node-by-node, choosing the parent-set that maximises each node's contribution to the LPL. The main problem with this class is that the composite model typically will not correspond to a single probability model. The output is therefore a simple heuristic. It is nevertheless very useful as an additional exploratory data analysis tool.

We call this approach the *MDM-DGM algorithm* (DGM is short for Directed Graph Model). The analysis of cyclic graphs is unlike the analysis of DAGs in some aspects. Spirtes, Glymour and Scheines (2000, Chapter 12) compare some properties such as the Markov condition and factorability between DAG and cyclic graphs. For instance, a DAG satisfies the local Markov property (LMP), *that is*, the variable of node i, given its parents, is independent of all other variables, except for its parents and descendants (Heckerman, 1998). But, a directed cyclic graph (DCG) does not always satisfy this property. Thus, suppose a DCG: $1 \to 2 \rightleftarrows 3$. Then although from one constrained graph: $1 \to 2 \to 3$, $\mathbf{Y}(3)$ would be independent of $\mathbf{Y}(1)$ given its parent $\mathbf{Y}(2)$ by the LMP. However, the subgraph $1 \to 2 \leftarrow 3$ implies that $\mathbf{Y}(3)$ is dependent of $\mathbf{Y}(1)$ given $\mathbf{Y}(2)$. So the LMP is not satisfied for this DCG.

Consider now another example, the DCG in Figure 3 where each variable is generated from another so that they are totally dependent. However, note that, in an associated undirected graph (changing all directed edges by undirected ones), $\mathbf{Y}(1)$ is conditional independent of $\mathbf{Y}(3)$ given $\mathbf{Y}(2)$ and $\mathbf{Y}(4)$ and, in the same way, $\mathbf{Y}(2)$ and $\mathbf{Y}(4)$ are conditional independent given $\mathbf{Y}(1)$ and $\mathbf{Y}(3)$. Therefore, Spirtes (1995) asserted that the notion of d-separation, while it does not imply conditional independence like in DAGs, is informative in cyclic graphs. Two disjoint sets of nodes, say **U** and **V**, are said to be *d-separated* by **W** if any element of **W** blocks every path between **U** and **V**. And a path is said to be *blocked* by a set of nodes, say **W**, if the path contains:

- a *chain*, $u \to y \to v$, where $y \in \mathbf{W}$, or

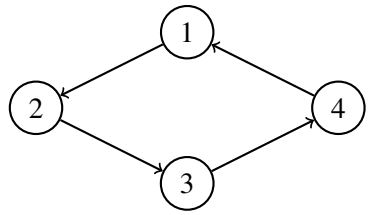

Figure 3 *An example of cyclic graph.*

- a *fork*, $u \leftarrow y \rightarrow v$, where $y \in \mathbf{W}$, or
- a *collider*, $u \rightarrow y \leftarrow v$, where $y \notin \mathbf{W}$ and no descendant of y is in $\mathbf{W}$.

In 2000, Spirtes et al. showed that the global Markov property for directed (acyclic or cyclic) graph holds for linear structural equation model (SEM). That is, a joint distribution P is represented by directed graph G if and only if whenever **I** and **II** are d-separated given **III** in G, where **I**, **II** and **III** are disjoint sets of nodes in G, the two sets of variables of nodes **I** and **II** are conditional independent in P given the variables of **III**.

Another aspect which differs between cyclic and acyclic graphs is factorability. In DAGs, the joint distribution of variables is defined as the product of the conditional distributions of each variable given its parents, whilst this may be not possible in DCG. For instance, for the 2-node DGM $1 \rightleftarrows 2$, a factorised distribution implies independence,

$$\begin{aligned} p\big(\mathbf{y}(1), \mathbf{y}(2)\big) &= p\big(\mathbf{y}(1)|\mathbf{y}(2)\big)p\big(\mathbf{y}(2)|\mathbf{y}(1)\big) \\ \Rightarrow \quad p\big(\mathbf{y}(1)|\mathbf{y}(2)\big) &= p\big(\mathbf{y}(1), \mathbf{y}(2)\big)/p\big(\mathbf{y}(2)|\mathbf{y}(1)\big) \\ &= p\big(\mathbf{y}(1)\big), \end{aligned}$$

which of course does not reflect the dependence constraint asserted by the graph. The joint model can no longer be guaranteed to be Gaussian and the calculation above can not provide a Bayesian conjugative analysis. In fact in the non-stochastic case the DGM-MDM degenerates to an SEM model, which are notoriously hard to formally estimate. In this sense, the DGM can be seen as a class of a structurally dynamic SEMs (see, *e.g.*, Koster, 1996). So this emphasises that the DGM models we fit here simply provide a heuristic, summarising the best features of classes of MDM and not a model itself.

Following, we show the performance of the search methods described here, using two real fMRI datasets. The first one consists of a resting-state experiment with 4 brain regions while the second one provides data from 5 different experiments with 11 brain regions.

4 The application of the MDM search methods into a resting-state fMRI real dataset

There is growing interest in the neuroscience literature about the brain at rest. In simple terms, the brain continues to work even when the person is apparently not performing any activity, and in particular, the brain spends about 20% of the bodies energy regardless of whether in a resting or task state (Raichle, 2010). Resting-state fMRI allows the study of intrinsic connectivity networks that represent a map of complex neural circuits. Initially, the resting-state data were measured with neuroimaging through positron emission topography (PET) and more recently through functional magnetic resonance imaging (fMRI). FMRI data reflect the blood oxygenation level, which is indirectly related to the activation of brain neurons. When neurons increase their firing rate, they require more oxygen, which in turn results in an increase in blood flow in that region. Counterintuitively, the amount of oxygen delivered by the blood exceeds the increased demand for oxygen. Therefore, an increase in oxygen is indicative of activation of the neurons in that place. This change in oxygenation gives rise to the blood oxygenation level dependent signal (BOLD), the time series variable measured by fMRI.

Our first application consists of a resting-state study in which participants were instructed to rest with their eyes open while the word "Relax" was centrally projected in white, against a black background (Shehzad et al., 2009, Ridgway et al., 2013). There are 25 right-handed native English-speaking participants, being 11 males with mean age of 20.5 ± 8.4. Subjects had no history of psychiatric or neurological illness, as confirmed by a psychiatric clinical assessment. They were scanned 3 times, being session 2 was $5 - 11$ months after the first, and session 3 was < 45 minutes after session 2. Data consist of 197 BOLD fMRI resting-state time-points, sampled every 2 seconds, for 4 ROI's: Posterior Cingulate—*PC*; Anterior Frontal—*AF*; Left Lateral Parietal—*LP* and Right Lateral Parietal—*RP*. According to Shehzad et al. (2009), "Mean time series for each ROI were extracted from this standardized functional volume by averaging over all voxels within the region. To ensure that each time series represented regionally specific neural activity, in each analysis, the mean time series of each ROI was orthogonalized with respect to 9 nuisance signals (global signal, white matter, cerebrospinal fluid, and 6 motion parameters)". These data are available for download in http://www.nitrc.org/projects/nyu_trt.

In this experiment, we are interested to study the information flow of the brain, and then we need to estimate the map of brain connectivity. First, four different graphical structures were manually chosen to representing the scientific beliefs about the brain connectivities (from DAG1 to DAG4 in the Figure 4). *DAG1* represents the idea that Posterior Cingulate hub drives other regions whilst *DAG2* corresponds to a Posterior Cingulate hub driven by Anterior Frontal and Left and Right Lateral Parietal. In *DAG3*, the information flows "forward" while, in *DAG4*,

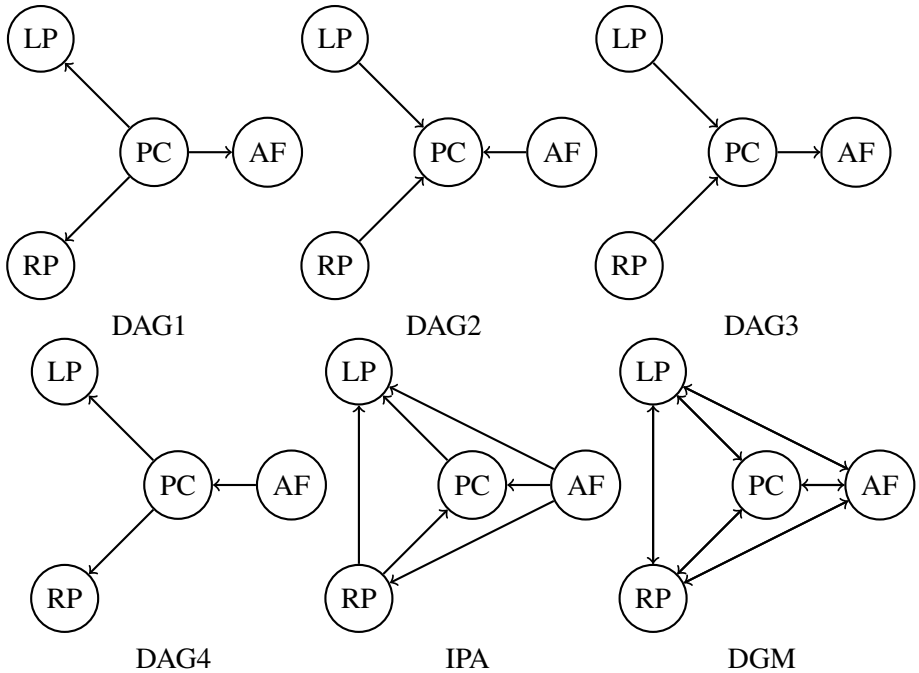

Figure 4 *The graphical structures from DAG1 to DAG4 were used in the first learning process for resting-state fMRI real data and DAG4 was chosen for most of datasets. Then the scores were summed over all datasets and the MDM-IPA and the MDM-DGM were applied. The results are IPA and DGM graphs, respectively. PC means the posterior cingulate area, AF means the anterior frontal area, LP means the left lateral parietal area and RP means the right lateral parietal area.*

the information flows "backward". Some previous studies with similar or with the same data (see, *e.g.*, Smith et al., 2009, Ridgway et al., 2013) which estimate functional connectivity have provided a graph formed by the union of these 4 DAGs, *that is*, there is a bidirectional edge between *LP* and *PC*, *RP* and *PC*, and *AF* and *PC*. And so, it was not possible to estimate the information flow as we can do below.

Considering a weakly informative priors, with $n_0(r) = d_0(r) = 0.001$ and $\mathbf{C}_0^*(r) = 3\mathbf{I}_{p_r}$ for all regions r, the discount factor was estimated for each region, for each model and each of the 75 datasets (see the average of δ across datasets in Figure 5). The regions that show the most dynamic behaviour (smaller δ) were *AF* and *PC*, but only in DAG models where these regions had no parents. This result is expected for such regions, in that, with no external input, variation observed must be explained by a highly variable internal state variable (*i.e.* intercept). The log predictive likelihood was used to select the best DAG model (of the four) for each session and each subject. DAG4 was selected a majority of datasets (54.7%), followed by DAG1 for 41.3% of runs. With respect to the intrasession consistency, we found that 91% of the subjects having a single consistent optimal DAG.

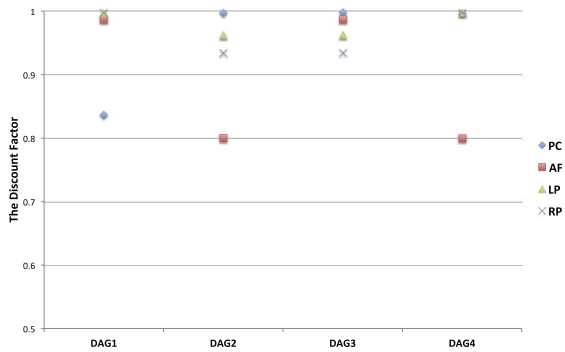

Figure 5 *The average of discount factor parameter δ across* 75 *datasets* (3 *sessions for each* 25 *subjects*) *for each DAG and region. PC means the posterior cingulate area, AF means the anterior frontal area, LP means the left lateral parietal area and RP means the right lateral parietal area.*

Markov equivalent and non-equivalent graphs

We focus on subject 22 and session 2, a typical where DAG4 was selected. Figure 6 shows the filtering (blue lines) and the smoothing (green lines) estimates for connectivities. Time-varying connectivity is evident: Connections out of the PC (to LP and RP) are stronger before time point 50 in comparison with the connections after time point 50 (Figures 6(a) and 6(b)). Figure 7 shows the LPL for different values of discount factor (note that here we use the same value of δ for all nodes). We consider three DAGs: DAG1 and DAG4 are *Markov equivalent* graphs while each is *Markov non-equivalent* to DAG3. The measure of model selection, logBF, is calculated as the difference between LPL of two DAGs, and so in Figure 7 the model with the highest curve for a given δ is optimal, and the highest point over all determines the best model. Therefore DAG4 should be chosen for all values of discount factor, except for $\delta = 1$ (the BN case) when the LPL is the same for DAG1 and DAG4, and it is not possible to distinguish between these two equivalent DAGs.

As a log Bayes factor is a difference of LPL scores, and the LPL is comprised of a sum of terms, one per time point, we can examine the evolution of the evidence for one model versus another with a plot of the cumulative log Bayes factor. For instance, Figure 8 shows the cumulative log Bayes factor comparing two Markov non-equivalent graphs, DAG4 with DAG3 (orange lines), using a static model (dotted lines) and dynamic models (solid lines). By about time point 40, evidence accumulates for DAG4 (static or dynamic) vs. DAG3, but by time point 100 the dynamic model is clearly favoured. Of course the two Markov equivalent graphs, DAG4 and DAG1 (blue lines), are indistinguishable in a static model (blue dotted line), while in the dynamic model (blue solid line), we see that evidence for DAG4 over DAG1 "arrives" by time 40 and is further bolstered around time 90. Therefore, analyses presented here show that the methods that estimate functional

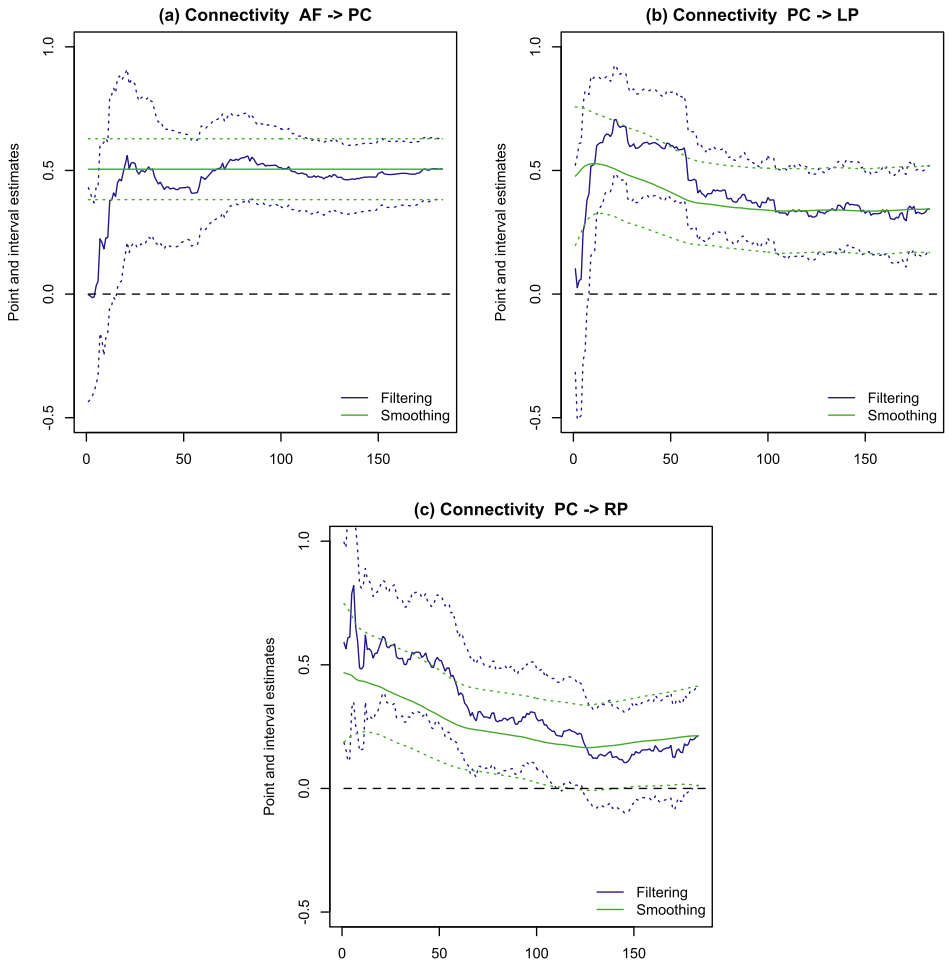

Figure 6 *The filtering (blue) and smoothing (green) posterior mean with 95% HPD interval for connectivities* (a) *AF* $\rightarrow$ *PC*, (b) *PC* $\rightarrow$ *LP*, (c) *PC* $\rightarrow$ *RP*.

connectivity, *e.g.* BN ($\delta = 1$), are not able to discriminate between DAG4 and DAG1 in this application, whilst the MDM that estimates effective connectivity provides the result that DAG4 was selected for most datasets.

Searching the MDM

In order to search all possible networks for each subject and session, we use the method detailed in Section 3. This requires that we compute the possible LPL contributions for each node, specifically the score for every possible set of parents. This score computation procedure took about 1 minute per dataset, using the software *R* (R Core Team, 2016) on a 2.7 GHz quad-core Intel Core i7 linux host with 16 GB. We summed individual scores over all datasets, to produce a

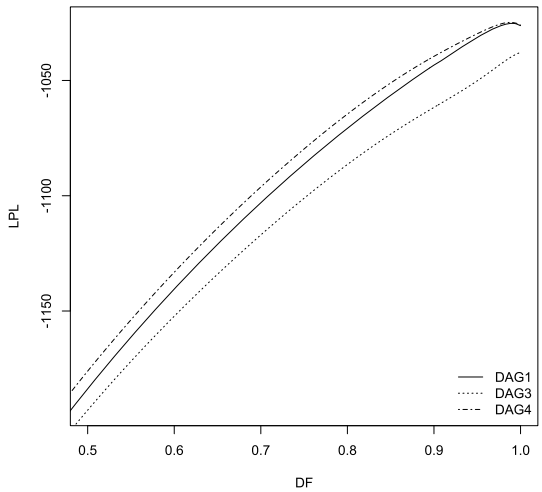

Figure 7 *The log predictive likelihood versus different values of discount factor* (*DF*). *DAG*1 (*solid line*) *and DAG*4 (*dotdashed line*) *are considered Markov equivalent whilst neither is equivalent to DAG*3 (*dotted line*).

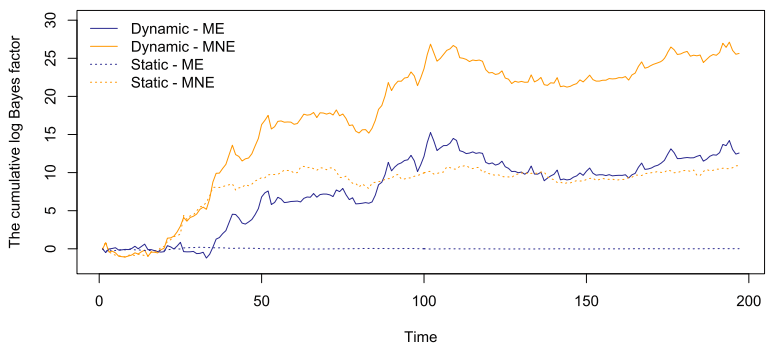

Figure 8 *The cumulative log Bayes factor comparing DAG*4 *to DAG*1, *Markov equivalent graphs* (*blue lines*), *and comparing DAG*4 *to DAG*3, *Markov non-equivalent graphs* (*orange lines*), *considering a static model* ($\delta = 1$; *dotted lines*) *and a dynamic model* ($\delta < 1$; *solid lines*).

model estimation procedure that identifies a single common network. Now, however, this does not constrain the connection strengths over subjects/sessions, just the graphical structure. The MDM-IPA was applied using GOBNILP (Globally Optimal Bayesian Network learning using Integer Linear Programming) which is a C program that learns networks from local scores using the SCIP framework for Constraint Integer Programming (Cussens, 2012; Bartlett and Cussens, 2013). The MDM-DGM was applied in the software R and both methods were run on a Intel 2.83 GHz Core2 Quad CPU with 8 GB RAM. The results were found almost instantly. The MDM-IPA procedure found a similar graph to DAG4 (Figure 4—IPA),

that is, the information flows in a "backward" way, except for the edge RP $\longrightarrow$ PC. However the MDM-DGM found causal interactions between all brain regions considered (see Figure 4—DGM).

5 The analysis of resting-state and task-based fMRI data

For the second application, we used a novel type of fMRI study that examined five separate five-minute steady-state sessions: Session 1 was a (conventional) resting-state condition; session 2 was a motor condition in which involved continuous and monotonic sequential finger tapping against the thumb, using the right hand; session 3 was a visual condition which consisted of videos of colourful abstract shapes in motion; session 4 and session 5 were a combination between visual and motor condition, but the former was in a random way whilst in the latter, individuals were instructed to change tapping direction when they saw an irregularly appearing cue, which were present in all visual conditions. Note the novel aspect, how all tasks are "on" for the entire acquisition, whereas conventional task-based fMRI entails alternating periods of stimulus and rest. Data were acquired on 15 healthy volunteers, and each acquisition consists of 230 time points, sampled every 1.3 seconds, with $2 \times 2 \times 2$ mm^3 voxels. The FSL software was used for preprocessing, including head motion correction, automated artifact removal procedure (Salimi-Khorshidi et al., 2014, Griffanti et al., 2014) and intersubject registration. The FSL tools FSLMaths and Randomise generate variance maps and variance change maps whilst FEAT and specialised code written in Python generate correlation maps (see details of data preprocessing in Duff et al., 2017). FSL (FMRIB Software Library) consists of analysis tools for fMRI, MRI and diffusion tensor imaging (DTI) brain imaging data. It uses Bayesian techniques to deal with the imperfect and noisy images of the brain and allows to incorporate prior belief about the brain and the neuroimaging equipment (Smith et al., 2004, Woolrich et al., 2009, Jenkinson et al., 2012).

We use 11 ROI's defined on 5 *motor* brain regions and 6 *visual* regions. The motor nodes are Cerebellum, Putamen, Supplementary Motor Area (SMA), Precentral Gyrus and Postcentral Gyrus (nodes numbered from 1 to 5 respectively) whilst the visual nodes are Visual Cortex V1, V2, V3, V4, V5 and task negative (V1 + V2; nodes numbered from 6 to 11 respectively). The observed time series are computed as the average of BOLD fMRI data over the voxels of each of these defined brain areas. See more details about these data in Duff et al. (2017). In this experiment, the estimates of the brain connectivity maps will contribute to scientific discussion about the difference between rest and task activity and how the communication between these brain regions varies across the five conditions of the fMRI dataset.

Searching the MDM

Both the MDM-IPA and the MDM-DGM require the scores for each subject, session and node, considering all possible sets of parents for each node. Computing these scores took about 5 minutes per subject/session, using the software *R* on a 2.7 GHz quad-core Intel Core i7 linux host with 16 GB. The MDM-IPA took less than one minute per subject/session, using GOBNILP, whilst the result of the MDM-DGM was found almost instantly in the software *R*, both on an Apple MacBook Pro with a Intel 2.83 GHz Core2 Quad CPU with 8 GB RAM. As the MDM-IPA and the MDM-DGM were applied in the same computer, we noted that the results for the latter were found a little faster. We assessed the intersubject consistency of the resulting networks by the prevalence of directed edges and by testing a null hypothesis of homogeneous connectivity over the network. Specifically, we estimated p_{ij}, the probability that an edge $i \to j$ exists, as the proportion $\hat{p}_{ij}$ of subjects with this particular edge. We used a one-sided Binomial test of $H_0 : p_{ij} = \pi$ versus $H_a : p_{ij} > \pi$, where π is the edge occurrence rate under homogeneity, set equal to the average of $\hat{p}_{ij}$ over the 90 possible edges.

Figure 9 shows $\hat{p}_{ij}$, but only for those edges with significant Binomial tests after false discovery rate correction (FDR; Benjamini and Hochberg, 1995) at level $\alpha_{\text{FDR}} = 0.05$, where i (parents) indexes rows and j (children) indexes columns (see Figure 17 for the unthresholded image of p_{ij}). The black horizontal and vertical lines divide the figure into four squares; the top left square represents the connectivity between motor brain regions, whilst the lower right square represents one between visual brain regions. Unsurprisingly, most of connectivities are within these two squares. The two other squares represent *cross-modal* connections, between motor and visual regions, that are less prevalent.

We applied the MDM-DGM algorithm across subjects as described previously. Significantly prevalent edges are shown in Figure 10, while all $\hat{p}_{ij}$ are shown in Figure 18. The nodes are ordered according to the expected flow of information in the brain, and thus it is notable that we find significant edges between consecutive nodes. In general, Figure 9 also shows this pattern but less clear for DAG constraints. Interesting patterns of *lack of edges* can be seen in Figure 18; for example, while cross-modal connections are inconsistent, with $\hat{p}$'s rarely exceeding 50%, efferent connections from primary motor and somatosensory (regions 4 and 5) to motor regions are particularly lower (appear in green) in all sessions. Also notable is how visual regions have inconsistent ($\approx 50\%$) influence on motor regions *except* for the visual session (session 3), where regions V1–V3 efferent connections to motor regions have notably lower prevalence.

Functional and effective connectivity

We also consider two methods of estimating the functional connectivity: *full correlation* and *partial correlation* (Baba, Shibata and Sibuya, 2004, Marrelec et al.,

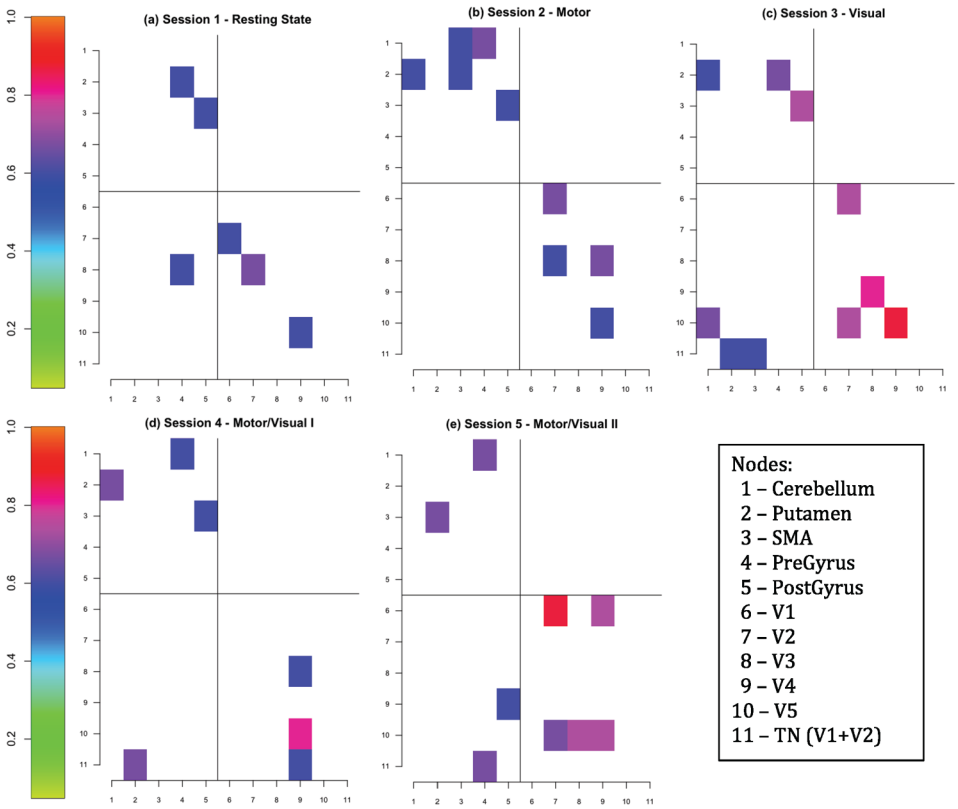

Figure 9 *The proportion of subjects who have a particular edge $i \to j$, where i indexes rows and j columns, using the* MDM-IPA *per session, only for significant connectivities,* $\alpha_{\text{FDR}} = 0.05$. *The black horizontal and vertical lines divide the figure into four squares; the top left square represents the connectivity between motor brain regions, whilst the lower right square represents one between visual brain regions. In general, intra-modal connections are more frequent than cross-modal connections.*

2006). For each node pair, per subject/session, we computed the full and partial correlation and converted each to a Z statistic with Fisher's transformation. For each node pair we tested the null hypothesis of mean zero (Fisher's transformed) correlation with a one-sample t-test, corrected for multiplicity with FDR ($\alpha_{\text{FDR}} = 0.05$). Figure 11 and Figure 12 show the significant ($\alpha_{\text{FDR}} = 0.05$) full and partial correlation for every session, respectively. Note that these techniques provide symmetric results about the principle diagonal. The vast majority of connections exist with high significance full correlation (Figure 11), however connections with the strongest correlation (above 0.6) tend to be intra-modal as discussed above. As expected, the significant MDM edges are a subset of the significant partial correlations (Figure 12).

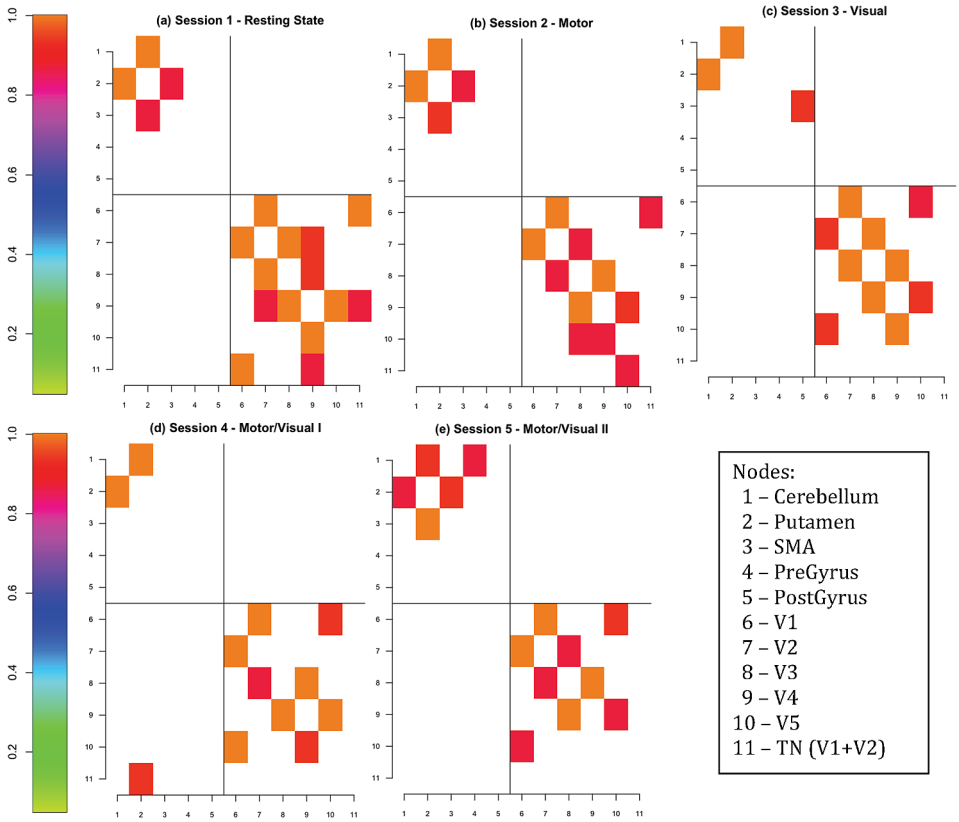

Figure 10 *The proportion of subjects who have a particular edge using the* MDM-DGM *per session, only for significant connectivities,* $\alpha_{\text{FDR}} = 0.05$ *(see Figure 9 for description of the panels). Within each group, nodes are arranged according to the anticipated flow of information in the brain and note the prevalence of forward and reverse connections along this hierarchy. Notably* $V1 \leftrightarrow$ *task-negative visual connections are found in resting, while* $V1 \leftrightarrow V5$ *connections are found in all tasks with a visual component (sessions 3–5).*

In short, while full and partial correlations do not account for nonstationarities nor represent a particular joint model, Figures 9 and 10 demonstrate that the application of the MDM gives scientifically plausible results and ones broadly compatible with other methods. For example, Duff et al. (2017) using these data and other functional connectivity analyses found that correlations between certain visual brain regions increased under visual stimulation. In this work, we have also shown that the strength of connectivity between some visual brain regions is higher in sessions that used visual stimulation than in other sessions, *for example*, in resting-state (comparing for example connectivities between visual regions in Figure 9(a) and (c)).

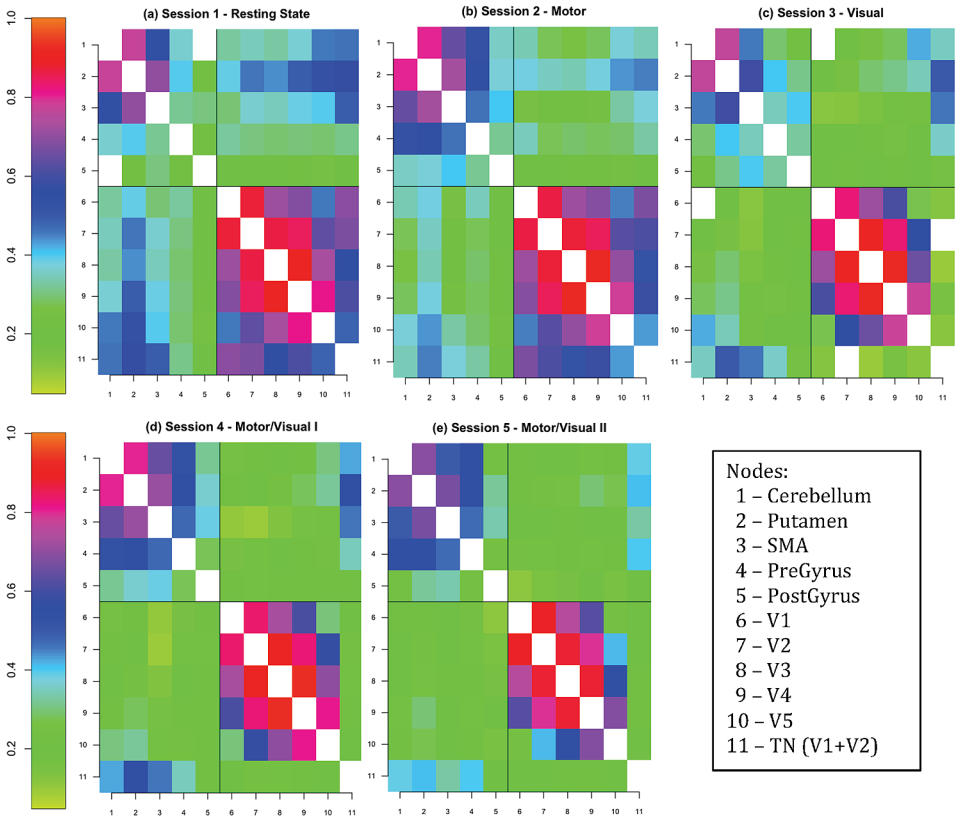

Figure 11 *The average correlation significant between two nodes over subjects using* full correlation method *for every session* ($\alpha_{\text{FDR}} = 0.05$). *Nearly all edges have significant non-zero correlations, as expected, and mostly within modality.*

Comparing sessions

We make formal comparisons between the different experimental conditions. For the MDM-IPA and the MDM-DGM, we compared edge prevalence between every possible pair of sessions using a McNemar test. Adjusting each of the $5 \times (5 - 1)/2 = 10$ possible session comparisons for FDR, there is no significant difference between the sessions for the MDM-IPA at FDR 5%. However, for the MDM-DGM algorithm, significant differences were found for 9 of 10 session comparisons, as shown in Figure 13. (Note that FDR controls the familywise error rate weakly, and thus less than one (*i.e.*, $10 \times 0.05 = 0.5$) of the comparisons are expected to show *any* positive results if all 10 were null; further, FDR control on each matrix is not compromised by examining the set of 9 results.) In general most of the difference between the sessions occurs in the connections between visual nodes (in the lower right square); in particular note how reciprocal V1–task-negative connections are more prevalent in resting-state relative to other sessions.

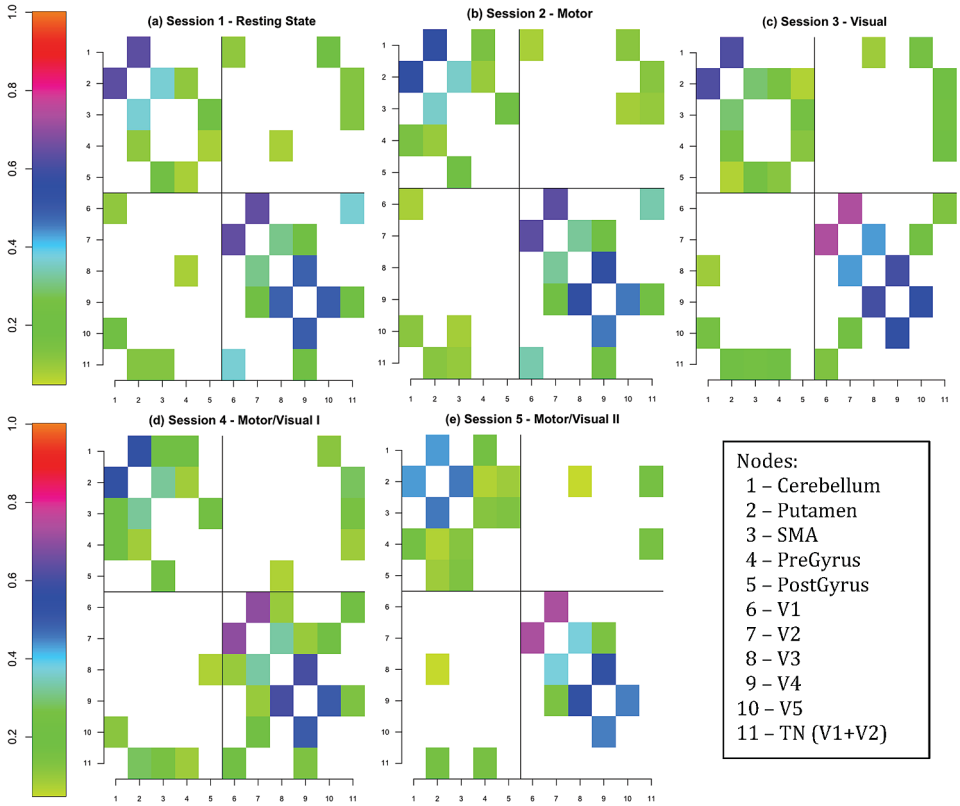

Figure 12 *The average correlation significant between two nodes over subjects using* partial correlation method *for every session* ($\alpha_{\text{FDR}} = 0.05$). *Partial correlations are seen to be consistent and along adjacent regions in the visual hierarchy* (*V*1–*V*5, *regions* 6–10) *and among the first* 3 *regions of the motor regions* (*cerebellum*, *putamen and SMA*).

When Session 1, resting-state, was compared with other sessions, most of the difference between connections was positive (with colours above the dark blue in the colour scale). So, in this case, connections existed in resting-state but not in other experimental conditions. Similar result was also found in Duff et al. (2017) who asserted that many correlations between nodes were lower during the active conditions compared to rest. Figures 14 and 15 gives the significant difference of full and partial correlations for every pair of sessions, respectively. Overall the number of significant different connections was highest between Session 1, resting-state, and other sessions. Session 3, visual condition, was closest to Session 4, visual and motor conditions.

Evaluating the dynamics of the system

As described above, model fitting involves estimating the discount factor δ for each node in each subject. Figure 16 shows the discount factor for each node (mean

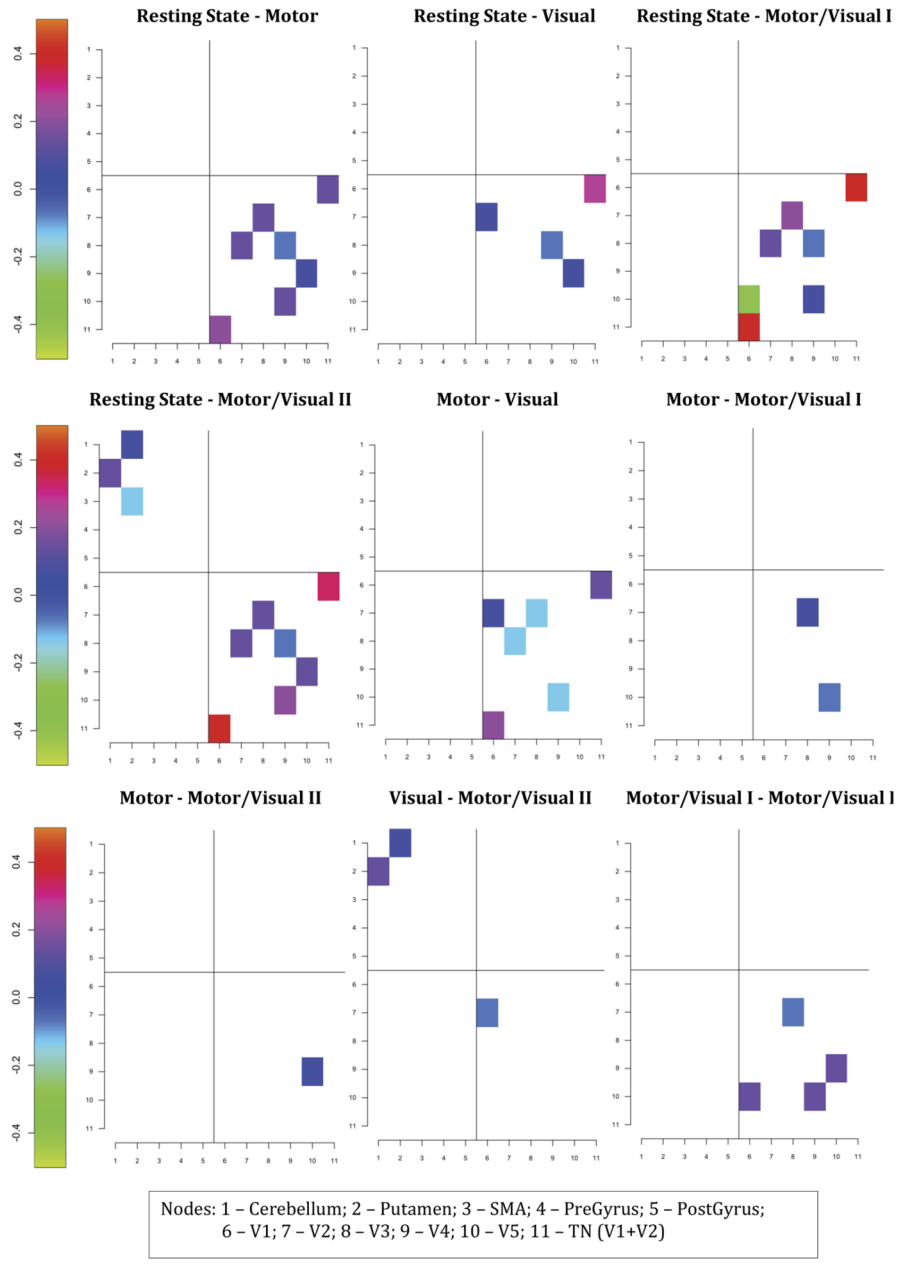

Figure 13 *The significant difference of the proportion of subjects who have a particular connection $i \to j$ between two sessions using the* MDM-DGM algorithm. *Positive intensities reflect greater edge prevalence in the first session listed over the second. For example, in the first (upper, left) panel comparing resting-state to the motor condition, pink, purple and dark blue indicate edges were found more often in resting session relative to the motor session. Prevalence never differed significantly in cross-modal areas.*

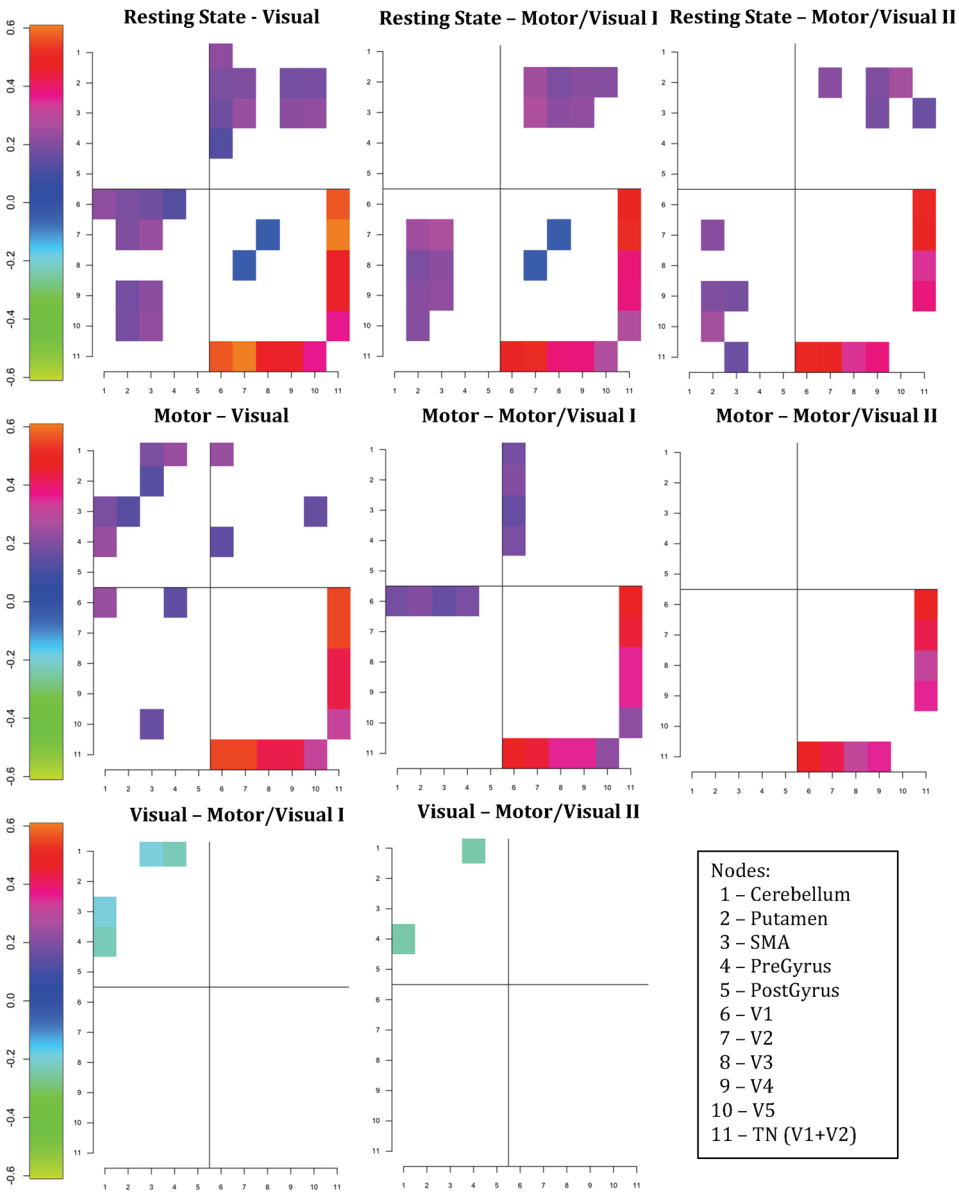

Figure 14 *The significant difference of the average of* full correlation *over subjects between two sessions. Differences in cross-model correlations are evident between resting* (*stronger*) *and other conditions* (*weaker*), *as well with the task-negative visual region* (*node* 11) *and all other visual regions* (*stronger in resting and motor sessions* 1 *and* 2).

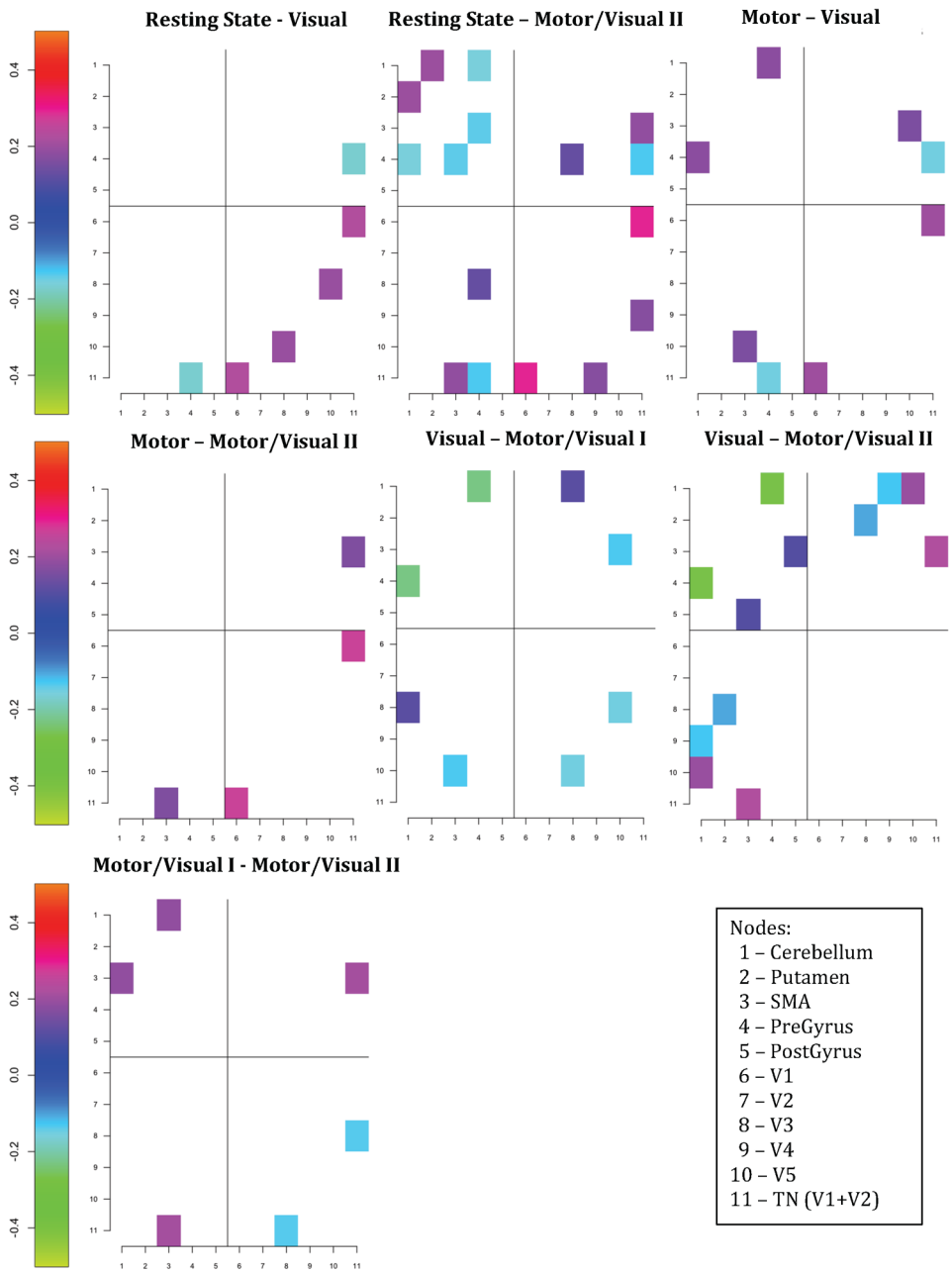

Figure 15 *The significant difference of the average of* partial correlation *over subjects between two sessions. Differences in connection strength are diffuse and difficult to interpret* (*compare with MDM-DGM differences in Figure* 13).

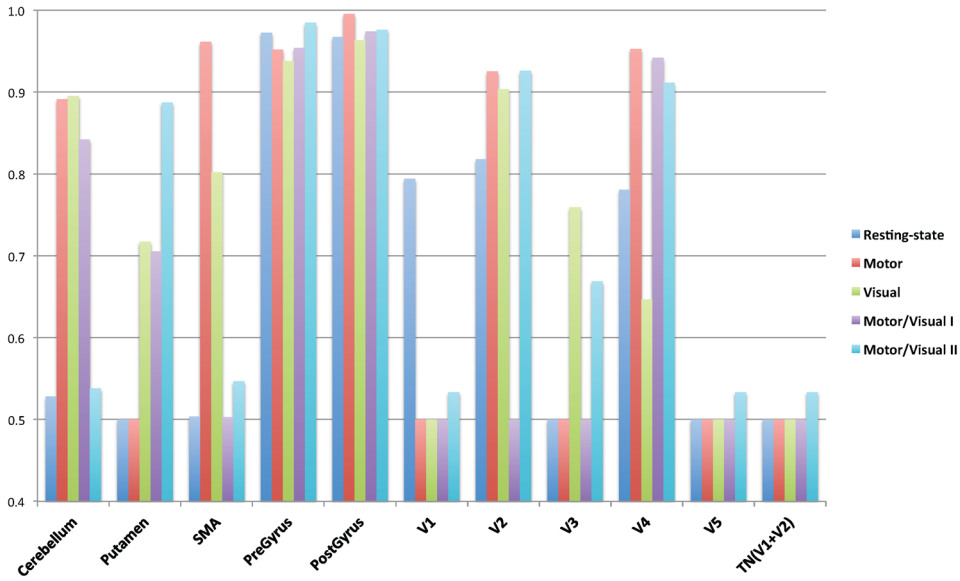

Figure 16 *The discount factor of each node averaged over the subjects using the graphical structure found in Figure 9 for each session. Precentral and postcentral gyrus are notable for having nearly stationary behaviour in all experiments, while V5 and task-negative visual regions are notable for having nearly maximal time-varying behavior.*

over subjects), for each session. For two motor areas, Precentral and Postcentral Gyrus (nodes 4 and 5) the average of δ is above 0.9 for all sessions, indicating a more static pattern. However, other brain areas appear to be noticeably more volatile. In general it appears that visual nodes have a shorter memory than motor nodes. For instance, the average of DF for nodes that have parents in Session 1 was 0.96 for motor nodes and was 0.80 for visual nodes. A possible reason is that the physical/sensory environment is much more constrained/static than the visual environment. For example, for session 1 (resting-state experiment), subjects were shown a screen with a fixation point. However, they were not explicitly asked to fixate. This might explain the greater perceptual variability in visual relative to sensory-motor areas.

Session 1, resting-state, has one of the smallest δ comparing with other sessions; an exception is Visual Cortex V1 (node 6) which, interestingly, has 1 or more parents in the resting session but none in the other sessions. One possible reason is that subjects do not do a specific activity in the resting-state experiment, being free to switch between different mental activities during the experiment. In contrast session 5, tapping cued by random events in the movie, has a long memory with some of the largest δ with the exception of Cerebellum, SMA and V4 areas (nodes 1, 3 and 9, respectively).

6 Conclusions

This paper presented the Multiregression Dynamic Model applied to fMRI data. Initially, we showed here, using the resting-state fMRI data, that the MDM can distinguish between Markov equivalent and non-equivalent DAGs whilst the BN can not do that. Therefore, the MDM estimates that the information in the brain flows "backwards" for most subjects. To our knowledge, this result has never been found before using this data. We also show for the first time an application of the MDM into task fMRI data.

Also we discussed two different search methods: The MDM-IPA and the MDM-DGM and presented their applications into neuroscience data. The analysis with fMRI real data using the MDM-IPA shows results consistent with neuroimaging literature. The estimated discount factor smaller than 1 shows that most connectivities vary over time so that there is a loss of information and a systematic bias when static models are used to study brain connections.

Although the MDM-IPA appears to fit fMRI data well, in order to obtain more realistic results, we used for the first time the MDM-DGM algorithm to learn directed (cyclic or acyclic) graphs. As shown above, the MDM-DGM provides graphs with most of its bidirectional connections between consecutive nodes (*i.e.* nodes anatomically closer to each other than to others). Another interesting result is that, as found in similar but independent studies, here the brain connections decrease from rest to steady-state conditions (see, *e.g.*, Duff et al., 2017). Moreover, the MDM-DGM usually provides denser graphs than the MDM-IPA. This should be expected because the space of possible directed graphs is higher than for DAGs with the same number of nodes. This complicates the interpretation somewhat. However the number of nodes can easily be penalised for example using a penalty function or by using non local priors.

A further important question relates to the heterogeneity among individuals. In this work we have fitted models independently to each subject, or constrained only the network structure to be the same (and then let all other parameters be fit independently over subjects). We are now investigating Bayesian hyperclustering techniques that find homogeneous sub-groups in terms of connectivity.

Appendix A: MDM equations

In this appendix, we provide details about the filtering and smoothing equations based on the Kalman filter (West and Harrison, 1997, Petris, Petrone and Campagnoli, 2009). Moreover, the one-step forecast conditional distributions is also described here (Queen and Albers, 2008).

Filtered distributions

The filtering densities are defined assuming firstly

$$\big(\boldsymbol{\theta}_{t-1}(r)|\mathbf{y}^{t-1},\phi(r)\big)\sim\mathcal{N}\big(\mathbf{m}_{t-1}(r),\mathbf{C}^*_{t-1}(r)\phi^{-1}(r)\big)\quad\text{and}\tag{A.1}$$

$$\big(\phi(r)|\mathbf{y}^{t-1}(r)\big)\sim\mathcal{G}\Big(\frac{n_{t-1}(r)}{2},\frac{d_{t-1}(r)}{2}\Big).\tag{A.2}$$

Thus, the marginal distribution of $\boldsymbol{\theta}_{t-1}(r)$ given the past is written as

$$\big(\boldsymbol{\theta}_{t-1}(r)|\mathbf{y}^{t-1}\big)\sim\mathcal{T}_{n_{t-1}(r)}\big(\mathbf{m}_{t-1}(r),\mathbf{C}_{t-1}(r)\big),\tag{A.3}$$

a noncentral t distribution with $n_{t-1}(r)$ degrees of freedom and parameters $\mathbf{m}_{t-1}(r)$ and $\mathbf{C}_{t-1}(r)=S_{t-1}(r)\mathbf{C}^*_{t-1}(r)$, where $S_{t-1}(r)=\frac{1}{\mathbb{E}[\phi(r)|\mathbf{y}^{t-1}(r)]}=\frac{d_{t-1}(r)}{n_{t-1}(r)}$.

By equations (2.1) and (A.1), the conditional prior distribution of $\boldsymbol{\theta}_t(r)$ given $\mathbf{y}^{t-1}(r)$ is

$$\big(\boldsymbol{\theta}_t(r)|\mathbf{y}^{t-1}(r),\phi(r)\big)\sim\mathcal{N}\big(\mathbf{m}_{t-1}(r),\mathbf{R}^*_t(r)\phi^{-1}(r)\big),\tag{A.4}$$

where $\mathbf{R}^*_t(r)=\mathbf{C}^*_{t-1}(r)+\mathbf{W}^*_t(r)$. Thus, from this result and by the observation equation, the conditional predictive distribution is

$$\big(Y_t(r)|\mathbf{y}^{t-1}(r),\phi(r)\big)\sim\mathcal{N}\big(f_t(r),Q^*_t(r)\phi^{-1}(r)\big),\tag{A.5}$$

where $f_t(r)=\mathbf{F}'_t(r)\mathbf{m}_{t-1}(r)$ and $Q^*_t(r)=\mathbf{F}'_t(r)\mathbf{R}^*_t(r)\mathbf{F}_t(r)+1$.

The conditional posterior distribution of $\boldsymbol{\theta}_t(r)$ given $\phi(r)$ is found through the property of multivariate Gaussian distribution. Thus, equations (A.4) and (A.5) can be combined to form the conditional (on $\phi(r)$) posterior of $\boldsymbol{\theta}_t(r)$ given all data up through time t:

$$\big(\boldsymbol{\theta}_t(r)|\mathbf{y}^t(r),\phi(r)\big)\sim\mathcal{N}\big(\mathbf{m}_t(r),\mathbf{C}^*_t(r)\phi^{-1}(r)\big),\tag{A.6}$$

where

$$\mathbf{m}_t(r)=\mathbf{m}_{t-1}(r)+\mathbf{R}^*_t(r)\mathbf{F}_t(r)\big(y_t(r)-\mathbf{F}'_t(r)\mathbf{m}_{t-1}(r)\big)/Q^*_t(r);\quad\text{and}$$

$$\mathbf{C}^*_t(r)\phi^{-1}(r)=\big(\mathbf{R}^*_t(r)-\mathbf{R}^*_t(r)\mathbf{F}_t(r)\mathbf{F}'_t(r)\mathbf{R}^*_t(r)/Q^*_t(r)\big)\phi^{-1}(r).$$

Now, using equations (A.2) and (A.5), the posterior distribution of ϕ is found as

$$\big(\phi(r)|\mathbf{y}^t(r)\big)\sim\mathcal{G}\Big(\frac{n_t(r)}{2},\frac{d_t(r)}{2}\Big),\tag{A.7}$$

where $n_t(r)=n_{t-1}(r)+1$ and $d_t(r)=d_{t-1}(r)+(y_t(r)-f_t(r))^2/Q^*_t(r)$.

The marginal posterior distribution of $\boldsymbol{\theta}_t(r)$ is found by equations (A.6) and (A.7), *that is*,

$$\big(\boldsymbol{\theta}_t(r)|\mathbf{y}^t(r)\big)\sim\mathcal{T}_{n_t(r)}\big(\mathbf{m}_t(r),\mathbf{C}_t(r)\big),$$

where $\mathbf{C}_t(r)=S_t(r)\mathbf{C}^*_t(r)$.

These results were found assuming the equations (A.1) and (A.2). However, these equations hold for all t, including $t = 0$.

This closed form of the recurrences assumes the innovation variance matrix, $\mathbf{W}_t^*$, is known. When it is not true, the different values of $\mathbf{W}_t^*$ are expressed in terms of the loss of information in the change between times $t-1$ and t. More precisely, the distribution of the innovation residual is

$$\big(\mathbf{w}_t(r)|\mathbf{y}^{t-1}(r), \phi(r)\big) \sim \mathcal{N}\big(\mathbf{0}, \mathbf{W}_t^*(r)\phi(r)^{-1}\big),$$

and so the prior distribution of $\boldsymbol{\theta}_t$ is

$$\big(\boldsymbol{\theta}_t(r)|\mathbf{y}^{t-1}(r), \phi(r)\big) \sim \mathcal{N}\big(\mathbf{m}_{t-1}(r), \mathbf{R}_t^*(r)\phi(r)^{-1}\big).$$

Thus, the prior variance of $\boldsymbol{\theta}_t$ (after observing $\mathbf{Y}_{t'}$, $t' < t$, but before seeing $\mathbf{Y}_t$) is written as function of the posterior variance at time $t-1$, $\mathbf{C}_{t-1}^*(r)$, plus a measure of the uncertainty, $\mathbf{W}_t^*(r)$. Therefore, the higher the value of $\mathbf{W}_t^*$, the lower the precision of information from $\boldsymbol{\theta}_{t-1}$ and the current observation y_t has more impact into estimating of $\boldsymbol{\theta}_t$ than the past observations $\mathbf{y}^{t-1}$. Now if we can assume that $\mathbf{R}_t^*(r) = \mathbf{C}_{t-1}^*(r) + \mathbf{W}_t^*(r)$ is well approximated by $\mathbf{C}_{t-1}^*/\delta$ for some $\delta \in (0, 1]$, we have a similar expression for

$$\mathbf{W}_t^* = \frac{1-\delta}{\delta}\mathbf{C}_{t-1}^*.$$

The discount factor δ is chosen as the value that maximizes the log predictive likelihood (LPL).

One-step conditional forecast distributions

The one-step forecast distribution can be found through the prior distribution of ϕ given the past (equation (A.2)) and the conditional distribution of Y_t given the past and the precision parameter ϕ (equation (A.5)), that is,

$$\big(Y_t(r)|\mathbf{y}^{t-1}\big) \sim \mathcal{T}_{n_{t-1}(r)}\big(f_t(r), Q_t(r)\big), \tag{A.8}$$

where $Q_t(r) = S_{t-1}(r)Q_t^*(r)$.

Smoothed distributions

The smoothing estimation follows retrospective analysis, starting with $t = T-1$ and continues until $t = 1$. First, the smoothing distributions of $\boldsymbol{\theta}_t(r)$ given the entire time series and the precision $\phi(r)$, for $t = 1, \ldots, T-1$, is written as:

$$\begin{aligned} &p\big(\boldsymbol{\theta}_t(r)|\mathbf{y}^T, \phi(r)\big) \\ &\quad = \int p\big(\boldsymbol{\theta}_t(r)|\boldsymbol{\theta}_{t+1}(r), \mathbf{y}^T, \phi(r)\big)\, p\big(\boldsymbol{\theta}_{t+1}(r)|\mathbf{y}^T, \phi(r)\big)\, d\boldsymbol{\theta}_{t+1}(r). \end{aligned} \tag{A.9}$$

Suppose first that the second integration term is

$$\big(\boldsymbol{\theta}_{t+1}(r)|\mathbf{y}^T, \phi(r)\big) \sim \mathcal{N}\big(\mathbf{sm}_{t+1}(r), \mathbf{sC}_{t+1}^*(r)\phi^{-1}(r)\big), \tag{A.10}$$

Using Bayes' theorem, the first integration term is

$$p\big(\boldsymbol{\theta}_t(r)|\boldsymbol{\theta}_{t+1}(r), \mathbf{y}^T, \phi(r)\big)$$
$$= \frac{p(\boldsymbol{\theta}_t(r)|\boldsymbol{\theta}_{t+1}(r), \mathbf{y}^t, \phi(r))p(\mathbf{y}_{t+1}, \ldots, \mathbf{y}_T|\boldsymbol{\theta}_t(r), \boldsymbol{\theta}_{t+1}(r), \mathbf{y}^t, \phi(r))}{p(\mathbf{y}_{t+1}, \ldots, \mathbf{y}_T|\boldsymbol{\theta}_{t+1}(r), \mathbf{y}^t, \phi(r))}.$$

But, $\mathbf{Y}_{t+1}, \ldots, \mathbf{Y}_T$ are independent of $\boldsymbol{\theta}_t(r)$ given $\boldsymbol{\theta}_{t+1}(r)$ (Queen and Smith, 1993) and thus $p(\boldsymbol{\theta}_t(r)|\boldsymbol{\theta}_{t+1}(r), \mathbf{y}^T, \phi(r)) = p(\boldsymbol{\theta}_t(r)|\boldsymbol{\theta}_{t+1}(r), \mathbf{y}^t, \phi(r))$. This is a gaussian distribution by equations (A.6) and (A.4) with parameters:

$$\begin{aligned}
&\mathbb{E}\big[\boldsymbol{\theta}_t(r)|\boldsymbol{\theta}_{t+1}(r), \mathbf{y}^t, \phi(r)\big] \\
&\quad = \mathbf{m}_t(r) + \mathbf{C}_t^*(r)\big(\mathbf{R}_{t+1}^*(r)\big)^{-1}\big(\boldsymbol{\theta}_{t+1}(r) - \mathbf{m}_t(r)\big); \\
&\mathrm{var}\big[\boldsymbol{\theta}_t(r)|\boldsymbol{\theta}_{t+1}(r), \mathbf{y}^t, \phi(r)\big] \\
&\quad = \big(\mathbf{C}_t^*(r) - \mathbf{C}_t^*(r)\big(\mathbf{R}_{t+1}^*(r)\big)^{-1}\mathbf{C}_t^*(r)\big)\phi^{-1}(r).
\end{aligned} \tag{A.11}$$

Returning to initial problem (equation (A.9)), the required density $p(\boldsymbol{\theta}_t(r)|\mathbf{y}^T, \phi(r))$ can be seen as the expectation value of $p(\boldsymbol{\theta}_t(r)|\boldsymbol{\theta}_{t+1}(r), \mathbf{y}^T, \phi(r))$ (equation (A.11)) with respect to $(\boldsymbol{\theta}_{t+1}(r)|\mathbf{y}^T, \phi(r))$ (equation (A.10)). Therefore, by the properties of the multivariate gaussian distribution, the conditional distribution of $\boldsymbol{\theta}_t(r)$ given $\mathbf{y}^T$ and ϕ is also gaussian with the following parameters:

$$\begin{aligned}
\mathbf{sm}_t(r) &= \mathbb{E}\big[\boldsymbol{\theta}_t(r)|\mathbf{y}^T, \phi(r)\big] \\
&= \mathbb{E}\big[\mathbb{E}\big(\boldsymbol{\theta}_t(r)|\boldsymbol{\theta}_{t+1}(r), \mathbf{y}^T, \phi(r)\big)|\mathbf{y}^T, \phi(r)\big] \\
&= \mathbf{m}_t(r) + \mathbf{C}_t^*(r)\big(\mathbf{R}_{t+1}^*(r)\big)^{-1}\big(\mathbf{sm}_{t+1}(r) - \mathbf{m}_t(r)\big); \\
\mathbf{sC}_t^*(r) &= \mathrm{var}\big[\boldsymbol{\theta}_t(r)|\mathbf{y}^T, \phi(r)\big] \\
&= \mathbb{E}\big[\mathrm{var}\big(\boldsymbol{\theta}_t(r)|\boldsymbol{\theta}_{t+1}(r), \mathbf{y}^T, \phi(r)\big)|\mathbf{y}^T, \phi(r)\big] \\
&\quad + \mathrm{var}\big[\mathbb{E}\big(\boldsymbol{\theta}_t(r)|\boldsymbol{\theta}_{t+1}(r), \mathbf{y}^T, \phi(r)\big)|\mathbf{y}^T, \phi(r)\big] \\
&= \big[\mathbf{C}_t^*(r) - \mathbf{C}_t^*(r)\big(\mathbf{R}_{t+1}^*(r)\big)^{-1}\big(\mathbf{R}_{t+1}^*(r) - \mathbf{sC}_{t+1}^*(r)\big) \\
&\quad \times \big(\mathbf{R}_{t+1}^*(r)\big)^{-1}\mathbf{C}_t^*(r)\big]\phi^{-1}(r)
\end{aligned}$$

Moreover, as the conditional distribution of $(\phi(r)|\mathbf{y}^T)$ is given by equation (A.7), then $(\boldsymbol{\theta}_t(r)|\mathbf{y}^T) \sim \mathcal{T}_{n_T(r)}(\mathbf{sm}_t(r), \mathbf{sC}_t(r))$, where $\mathbf{sC}_t(r) = S_T(r)\mathbf{sC}_t^*(r)$.

The equation (A.10) is true for $t = T - 1$, that is

$$\big(\boldsymbol{\theta}_T(r)|\mathbf{y}^T, \phi(r)\big) \sim \mathcal{N}\big(\mathbf{sm}_T(r) = \mathbf{m}_T(r), \mathbf{sC}_{t+1}^*(r)\phi^{-1}(r) = \mathbf{C}_T^*(r)\phi^{-1}(r)\big).$$

Therefore, the distributions of $(\boldsymbol{\theta}_t(r)|\mathbf{y}^T)$ for $t = T - 1, T - 2, \ldots, 1$ can be computed by backward procedure.

Appendix B: Learning network results

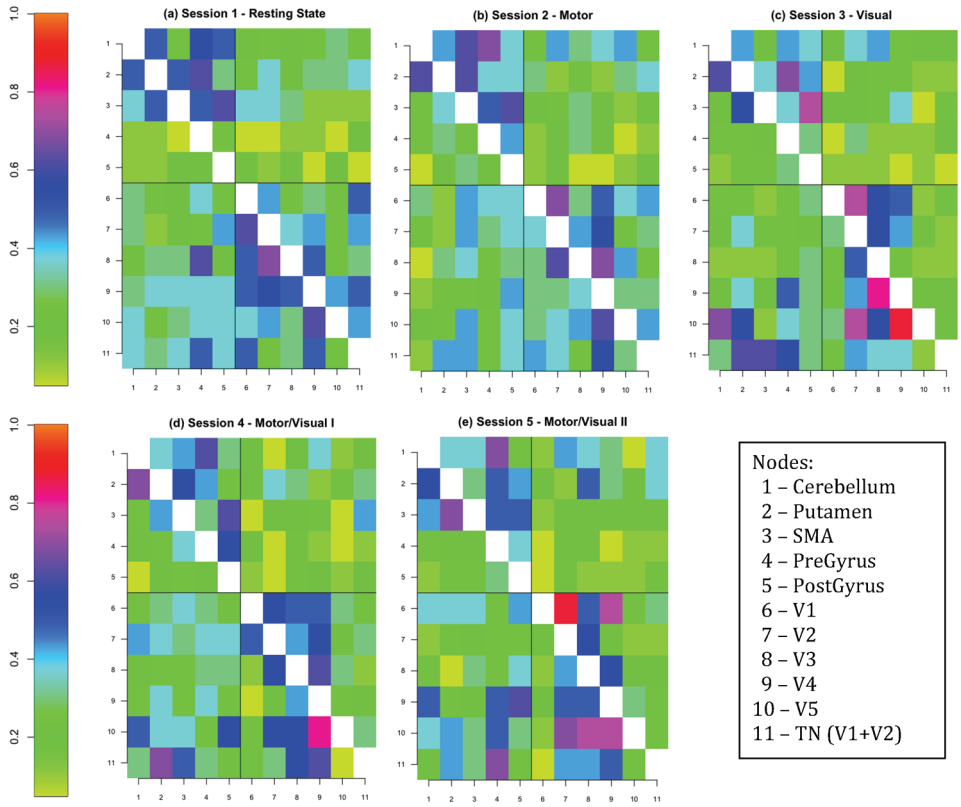

Figure 17 *The proportion of subjects who have a particular edge using the* MDM-IPA *per session* (*see Figure* 9 *for explanation of the panels*). *Clear intra- vs inter-modal pattern is seen*, *but a distinct pattern of hierarchical connections* (*e.g. among visual regions*) *is not seen*; *compare to Figure* 18.

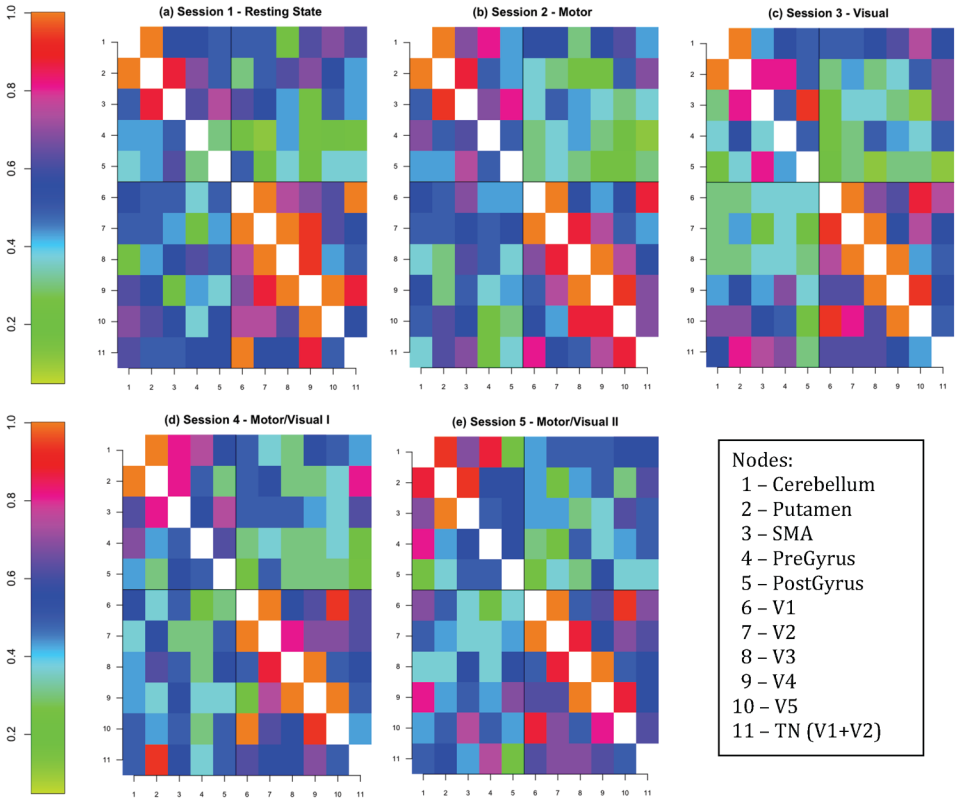

Figure 18 *The proportion of subjects who have a particular edge using the* MDM-DGM *per session* (*see Figure* 9 *for an explanation of the panels*). *A pattern of reciprocal adjacent connections up- and down the visual hierarchy* (*notes* 6–10, *V*1–*V*5) *is evident*.

Acknowledgments

This work was supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1, by CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior), Brazil and the Wellcome Trust. We are grateful to Tamar Makin and Eugene Duff for use of the steady state fMRI data.

References

Achterberg, T. (2007). Constraint integer programming. PhD thesis, TU Berlin.

Baba, K., Shibata, R. and Sibuya, M. (2004). Partial correlation and conditional correlation as measures of conditional independence. *Australian and New Zealand Journal of Statistics* **46**, 4, 657–664.

Bartlett, M. and Cussens, J. (2013). Advances in Bayesian network learning using integer programming. arXiv preprint. Available at arXiv:1309.6825.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* **57**, 1, 289–300.

Chang, C., Thomason, M. E. and Glover, G. H. (2008). Mapping and correction of vascular hemodynamic latency in the BOLD signal. *NeuroImage* **43**, 90–102.

Costa, L., Smith, J., Nichols, T., Cussens, J., Duff, E. P. and Makin, T. R. (2015). Searching multiregression dynamic models of resting-state fMRI networks using integer programming. *Bayesian Analysis* **10**, 441–478.

Cowell, R. G. (2013). A simple greedy algorithm for reconstructing pedigrees. *Theoretical Population Biology* **83**, 55–63.

Cussens, J. (2010). SMaximum likelihood pedigree reconstruction using integer programming. *WCB@ ICLP* 8–19.

Cussens, J. (2012). Bayesian network learning with cutting planes. arXiv preprint. Available at arXiv:1202.3713.

David, O., Guillemain, I., Saillet, S., Reyt, S., Deransart, C., Segebarth, C. and Depaulis, A. (2008). Identifying neural drivers with functional MRI: An electrophysiological validation. *PLoS Biology* **6**, 2683–2697.

Duff, E., Tamar, M., Smith, S. M. and Woolrich, M. W. (2017). Disambiguating brain functional connectivity. bioRxiv. http://biorxiv.org/content/early/2017/01/25/103002.

Friston, K. J. (2011). Functional and Effective Connectivity: a review. *Brain Connectivity* **1**, 1, 13–36.

Friston, K. J., Harrison, L. and Penny, W. (2003). Dynamic causal modelling. *NeuroImage* **19**, 1273–1302.

Goldenberg, A., Zheng, A. X., Fienberg, S. E., Airoldi, E. M. and others (2010). A survey of statistical network models. *Foundations and Trends® in Machine Learning* **2**, 2, 129–233.

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* **37**, 424–438.

Griffanti, L., Salimi-Khorshidi, G., Beckmann, C. F., Auerbach, E. J., Douaud, G., Sexton, C. E., Zsoldos, E., Ebmeier, K. P., Filippini, N., Mackay, C. E. and Moeller, S. (2014). ICA-based artefact removal and accelerated fMRI acquisition for improved resting state network imaging. *NeuroImage* **95**, 232–247.

Havlicek, M., Jan, J., Brazdil, M. and Calhoun, V. D. (2010). Dynamic Granger causality based on Kalman filter for evaluation of functional network connectivity in fMRI data. *NeuroImage* **53**, 65–77.

Heckerman, D. (1998). A tutorial on learning with Bayesian networks. *Nato Asi Series D Behavioural And Social Sciences* **89**, 301–354.

Jeffreys, H. (1961). *Theory of Probability*, 3rd ed. London: Oxford University Press.

Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W. and Smith, S. M. (2012). FSL. *NeuroImage* **62**, 782–790.

Koster, J. T. (1996). Markov properties of nonrecursive causal models. *The Annals of Statistics* 2148–2177.

Marrelec, G., Krainik, A., Duffau, H., Pélégrini-Issac, M., Lehéricy, S., Doyon, J. and Benali, H. (2006). Partial correlation for functional brain interactivity investigation in functional MRI. *NeuroImage* **62**, 228–237.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.

Penny, W., Ghahramani, Z. and Friston, K. (2005). Bilinear dynamical systems. *Philosophical Transactions of the Royal Society of London Series B, Biological Sciences* **360**, 983–993.

Petris, G., Petrone, S. and Campagnoli, P. (2009). *Dynamic Linear Models with R*. New York: Springer.

Poldrack, R. A., Mumford, J. A. and Nichols, T. E. (2011). *Handbook of fMRI Data Analysis*. Cambridge University Press.

Queen, C. M. and Albers, C. J. (2008). Forecast covariances in the linear multiregression dynamic model. *J. Forecast.* **27**, 175–191.

Queen, C. M. and Albers, C. J. (2009). Intervention and causality: Forecasting traffic flows using a dynamic Bayesian network. *Journal of the American Statistical Association* **104**, 669–681.

Queen, C. M. and Smith, J. Q. (1993). Multiregression dynamic models. *Journal of the Royal Statistical Society, Series B* **55**, 849–870.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Raichle, M. E. (2010). Two views of brain function. *Trends in Cognitive Sciences* **14**, 180–190.

Ramsey, J. D., Hanson, S. J., Hanson, C., Halchenko, Y. O., Poldrack, R. A. and Glymour, C. (2010). Six problems for causal inference from fMRI. *NeuroImage* **49**, 1545–1558.

Ridgway, G., Leite, A. B., Penny, W. and Friston, K. (2013). Stochastic DCM of the DMN using resting-state fMRI: test-retest reliability. figshare. https://doi.org/10.6084/m9.figshare.866771.v1.

Ryali, S., Supekar, K., Chen, T. and Menon, V. (2011). Multivariate dynamical systems models for estimating causal interactions in fMRI. *NeuroImage* **54**, 807–823.

Salimi-Khorshidi, G., Douaud, G., Beckmann, C. F., Glasser, M. F., Griffanti, L. and Smith, S. M. (2014). Automatic denoising of functional MRI data: Combining independent component analysis and hierarchical fusion of classifiers. *NeuroImage* **90**, 449–468.

Schwab, S., Harbord, R., Costa, L. and Nichols, T. E. (2017). multdyn: A package for Multiregression Dynamic Models (MDM). Available at https://github.com/schw4b/multdyn.

Shehzad, Z., Kelly, A. C., Reiss, P. T., Gee, D. G., Gotimer, K., Uddin, L. Q., Lee, S. H., Margulies, D. S., Roy, A. K., Biswal, B. B. and Petkova, E. (2009). The resting brain: Unconstrained yet reliable. *Cerebral Cortex* **19**, 2209–2229.

Sloane, N. J. A. and Plouffe, S. (1995). *The Encyclopedia of Integer Sequences*. Academic Press.

Smith, J. F., Pillai, A., Chen, K. and Horwitz, B. (2010). Identification and validation of effective connectivity networks in functional magnetic resonance imaging using switching linear dynamic systems. *NeuroImage* **52**, 1027–1040.

Smith, J. F., Pillai, A., Chen, K. and Horwitz, B. (2011). Effective connectivity modeling for fMRI: Six issues and possible solutions using linear dynamic systems. *Frontiers in Systems Neuroscience* **5**, 104.

Smith, J. Q. and Croft, J. (2003). Bayesian networks for discrete multivariate data: An algebraic approach to inference. *Journal of Multivariate Analysis* **84**, 387–402.

Smith, S. M., Bandettini, P. A., Miller, K. L., Behrens, T. E. J., Friston, K. J., David, O., Liue, T., Woolricha, M. W. and Nichols, T. E. (2012). The danger of systematic bias in group-level FMRI-lag-based causality estimation. *NeuroImage* **59**, 1228–1229.

Smith, S. M., Fox, P. T., Miller, K. L., Glahn, D. C., Fox, P. M., Mackay, C. E., Filippini, N., Watkins, K. E., Toro, R., Laird, A. R. and Beckmann, C. F. (2009). Correspondence of the brain's functional architecture during activation and rest. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 13040–13045.

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E. J., Johansen-Berg, H., Bannister, P. R., De Luca, M., Drobnjak, I., Flitney, D. E., Niazy, R., Saunders, J., Vickers, J., Zhang, Y., De Stefano, N., Brady, J. M. and Matthews, P. M. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage* **23**, 208–219.

Spirtes, P. (1995). Directed cyclic graphical representations of feedback models. In *Uncertainty in Artificial Intelligence* **11** (P. Besnard and S. Hanks, eds.) 491–498. Morgan Kaufmann.

Spirtes, P., Glymour, C. N. and Scheines, R. (2000). *Causation, Prediction, and Search*, 2nd ed. Cambridge, MA: MIT Press.

Sporns, O. (2010). *Networks of the Brain*, 1st ed. MIT Press.

Stephan, K. E., Kasper, L., Harrison, L. M., Daunizeau, J., den Ouden, H. E., Breakspear, M. and Friston, K. J. (2008). Nonlinear dynamic causal models for fMRI. *NeuroImage* **42**, 649–662.

Valdés-Sosa, P. A., Roebroeck, A., Daunizeau, J. and Friston, K. (2011). Effective connectivity: Influence, causality and biophysical modeling. *NeuroImage* **58**, 339–361.

West, M. and Harrison, P. J. (1997). *Bayesian Forecasting and Dynamic Models*, 2nd ed. New York: Springer.

Williams, H. P. (2009). *Logic and Integer Programming*. Springer. ISBN 978-0-387-92279-9.

Woolrich, M. W., Jbabdi, S., Patenaude, B., Chappell, M., Makni, S., Behrens, T., Beckmann, C., Jenkinson, M. and Smith, S. M. (2009). Bayesian analysis of neuroimaging data in FSL. *NeuroImage* **45**, S173–S186.

L. Costa
Instituto de Matemática e Estatística
UFBA
Av. Adhemar de Barros, S/N, Ondina
CEP: 40170-110—Salvador/BA
Brasil
E-mail: liliacosta@ufba.br

T. Nichols
Big Data Institute
University of Oxford
Li Ka Shing Centre
for Health Information and Discovery
Old Road Campus
Oxford OX3 7LF
UK
E-mail: thomas.nichols@bdi.ox.ac.uk

J. Q. Smith
Department of Statistics
University of Warwick
Coventry, CV4 7AL
UK
E-mail: stran@live.warwick.ac.uk